**Department of Mathematics and Statistics**

# Modelling the Structure of Australian Wool Auction Prices

**Chi Ngok Chow**

**This thesis is presented for the Degree of**
**Doctor of Philosophy**
**of**
**Curtin University**

**September 2010**

**DECLARATION**

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature: ……..

Date: ……...

# ABSTRACT

The largest wool exporter in the world is Australia, where wool being a major export is worth over AUD $2 billion per year and constitutes about 17 per cent of all agricultural exports. Most Australian wool is sold by auctions in three regional centres. The prices paid in these auction markets are used by the Australian production and service sectors to identify the quality preferences of the international retail markets and the intermediate processors. One ongoing problem faced by wool growers has been the lack of clear market signals on the relative importance of wool attributes with respect to the price they receive at auction. The goal of our research is to model the structure of Australian wool auction prices. We aim to optimise the information that can be extracted and used by the production and service sectors in producing and distributing the raw wool clip.

Most of the previous methods of modelling and predicting wool auction prices employed by the industry have involved multiple-linear regressions. These methods have proven to be inadequate because they have too many assumptions and deficiencies. This has prompted alternative approaches such as neural networks and tree-based regression methods. In this thesis we discuss these alternative approaches. We observe that neural network methods offer good prediction accuracy of price but give minimal understanding of the price driving variables. On the other hand, tree-based regression methods offer good interpretability of the price driving characteristics but do not give good prediction accuracy of price. This motivates a hybrid approach that combines the best of the tree-based methods and neural networks, offering both prediction accuracy and interpretability.

Additionally, there also exists a wool specifications problem. Industrial sorting of wool during harvest, and at the start of processing, assembles wool in bins according to the required wool specifications. At present this assembly is done by constraining the range of all specifications in each bin, and having either a very large number of bins, or a large variance of characteristics within each bin.

Multiple-linear regression on price does not provide additional useful information that would streamline this process, nor does it assist in delineating the specifications of individual bins.

In this thesis we will present a hybrid modular approach combining the interpretability of a regression tree with the prediction accuracy of neural networks. Our procedure was inspired by Breiman and Shang's idea of a "representer tree" (also known as a "born again tree") but with two main modifications: 1) we use a much more accurate Neural Network in place of a multiple tree method, and 2) we use our own modified smearing method which involves adding Gaussian noise. Our methodology has not previously been used for wool auction data and the accompanying price prediction problem. The numeric predictions from our method are highly competitive with other methods. Our method also provides an unprecedented level of clarity and interpretability of the price driving variables in the form of tree diagrams, and the tabular form of these trees developed in our research. These are extremely useful for wool growers and other casual observers who may not have a higher level understanding of modelling and mathematics. This method is also highly modular and can be continually extended and improved. We will detail this approach and illustrate it with real data.

The more accurate modelling and analysis helps wool growers to better understand the market behaviour. If the important factors are identified, then effective strategies can be developed to maximise return to the growers.

In Chapter 1 of this thesis, we present a brief overview of the Australian wool auction market. We then discuss the problems faced by the wool growers and their significance, which motivate our research.

In Chapter 2, we define the predictive aspect of the modelling problem and present the data that is available to us for our research. We introduce the assumptions that must be made in order to model the auction data and predict the wool prices.

Chapter 3 discusses neural networks and their potential in our wool auction problem. Neural networks are known to give good results in many modern applications resolving industrial problems. As a result of the popularity of such methods and the ongoing development of them, our research partner, the Department of Agriculture and Food, Government of Western Australia, performed a preliminary investigation into neural networks and found them to give satisfactory predictions of wool auction prices. In our Chapter 3, we perform an analysis and assessment of neural networks, specifically, the generalised regression neural networks (GRNN). We look at the strengths and weaknesses of GRNN, and apply them to the wool auction problem and comment on their relevance and usability in our wool problem. We detail the problems we face, and why neural networks alone may not be the best approach for the wool auction problem, thus laying the foundation for the development of our hybrid modular approach in Chapter 5. We also use the numerical prediction results from GRNN as the benchmark in our comparisons of different modelling methods in the rest of this thesis.

Chapter 4 details the tree-based regression methods, as an alternate approach to neural networks. In analysing the tree-based methods with our wool auction data, we illustrate the tree methods' advantages over neural networks, as well as the trade-offs, with our auction data. We also demonstrate how powerful and useful a tree diagram can be to the wool auction problem. And in this Chapter, we improve a typical tree diagram further by introducing our own tabular form of the tree, which can be of immerse use to wool growers. In particular, we can use our tabular form to solve the wool specification problem mentioned earlier, and we incorporate this tabular form as part of a new hybrid methodology in Chapter 5. In Chapter 4 we also consider the ensemble methods such as bootstrap aggregating (bagging) and random forests, and discuss their results. We demonstrate that, the ensemble methods provide higher prediction accuracies than ordinary regression trees by introducing many trees into the model. But this is at the expense of losing the simplicity and clarity of having only a single tree. However, the study of assemble methods do end up providing an excellent idea for our hybrid approach in Chapter 5.

Chapter 5 details the new hybrid approach we developed as a result of our work in Chapters 3 and 4 using neural networks and tree-based regression methods. Our hybrid approach combines the two methods with their respective strengths. We apply our new approach to the data, compare the results with our earlier work in neural networks and tree-based regression methods, then discuss the results.

Finally, we conclude our thesis with Chapter 6, discussing the potential of our new hybrid approach and the directions of possible future works.

# ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor for this project, Prof. Lou Caccetta, for all his guidance and putting up with my various problems in the past couple of years. Thank you Lou!

I would like to thank my associate supervisor, A/Prof. John Stanton, for sharing his vast knowledge of the Australian wool industry and helping me understand what I needed for this project. I would also like to thank his friend and colleague from the Department of Agriculture and Food, WA, Dr Kimbal Curtis, for introducing me to NeuralTools and sharing interesting ideas which were very useful in the development of my hybrid method.

I would like to thank Dr Tony Dixon during his tenure at WACEIO, in helping to mentor me and introducing me to coding in R. Thank you Dr Ritu Gupta, for suggesting me to look up Bagging. I would like to thank Dr Mark Grigoleit for helping me set up much needed remote access on the computers I used and for fixing the network issues I faced.

I would like to thank my best friends in university, David Belton and Ian van Loosen. Thank you David for being such a good friend since our undergraduate days, and for your always helpful advices and encouragements during lunch breaks. And thank you for helping me code in Matlab! Thank you Ian for being a fellow fanboy, and for providing the Maths department with much comedy relief.

And I would like to thank all of my teachers, mentors, friends and colleagues from the Department of Mathematics and Statistics at Curtin University. I would also like to thank all friends and loved ones outside my academic life who supported me. Most importantly, I thank my parents, grandparents and sister for their support. You are the best family one can ever have.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF PUBLICATIONS RELATED TO THIS THESIS

Cheng Y.W., J. Stanton and L. Caccetta , Predicting the Australian Wool Price by Tree-based Regression, in Industrial Optimization Symposium 2003, (Caccetta L. and Rehbock V., Editors), Volume 1 (2004), 115-124.

Caccetta L., C. Chow, T. Dixon, and J. Stanton (2005), Modelling the Structure of Australian Wool Auction Prices, Conference Proceedings, International Congress on Modelling and Simulation (MODSIM05), (Editors: A. Zerger and R.M. Argent), Melbourne, Australia, December, 2005.
http://www.mssanz.org.au/modsim05/papers/caccetta_1.pdf

Caccetta L., C.N. Chow, K. Curtis, and J. Stanton (2007), Modelling the Structure of Australian Wool Auction Prices: A Hybrid Approach Combining Regression Tree and Neural Networks, Proceedings, The 7th International Conference on Optimization: Techniques and Applications (ICOTA7), (Editors: Masao Fukushima et al.), Kobe, Japan, December, 2007. ISBN 978-4-946443-15-2.

Caccetta L., C.N. Chow, and J. Stanton (2009), Modelling the Structure of Australian Wool Auction Prices: A Hybrid Approach Combining Regression Tree and Neural Networks, Conference Proceedings, The 20th National Conference of Australian Society for Operations Research (ASOR Conference 2009), Gold Coast, Australia, September, 2009.

# Chapter 1

# Introduction

Australian wool auction, worth over $2 billion per year, is an on-going process. The prices paid in this auction market are used by the Australian production and service sectors to identify the quality preferences the international retail markets and the intermediate processors. The aim of our research is to optimise the information that can be extracted and used by these sectors in the production and distribution of the raw wool clip.

In this introductory chapter we will present a brief overview of the Australian wool auction market. We will then discuss the problems faced by the wool growers and their significance, which motivate our research. At the end of this chapter we will present an overview of the rest of our thesis.

## 1.1 Australian Wool

70% of world trade in apparel wool is Australian wool. Unlike other commodities, each farm lot of wool is laboratory tested for its measurements/specifications, and each farm lot has an individual price. About 450,000 farm lots of wool are sold in Australia raw wool auctions each year. Raw wool is one of Australia's largest export commodities and is worth over AUD $2 billion annually. It constitutes around 17 per cent of all farm exports in Australia. Hence, Australia continues to be the largest exporter of wool in the world.

The auction centres for wool in Australia are located in three regions: Northern (Sydney, Newcastle, Goulburn), Southern (Melbourne, Geelong, Adelaide, Launceston) and Western (Fremantle). Auctions are held almost every week in the year, with breaks during the Easter period, a period in July, and the Christmas

period. Depending on the volume of wool sale lots in each auction, the auction can be spread over 1 to 3 days.


Figure 1.1: Auction Centres for Wool in Australia

The auctions are conducted using the greasy price for the raw wool, expressed in cents per kilogram. This can be converted to a clean price estimate by multiplying greasy price by 100 and dividing by the yield. The response variable is the clean price, which is the base price less the total discount in c/kg clean. The base price for wool for a given style and fibre diameter (micron), assuming there are no faults for strength, length, vegetable matter or colour is expressed in c/kg clean.

## 1.2 Wool Prices and Market Reporting

Before being put up for auction, each bale or sale lot of wool is submitted for third party laboratory testing, measuring the wool quality characteristics. This information is then made available to potential buyers for their considerations. The laboratory testing procedure is paid for by the wool grower and he/she can select which attributes are to be reported or withheld to maximise the

attractiveness of the sale lot to be auctioned. The most typical wool quality characteristics to be laboratory tested that are well accepted by the industry to be significant price-driving variables are: fibre diameter, proportion of break (middle), staple length, staple strength, vegetable matter content, and yield. They are described in Table 1.1 below. Of these, it is commonly accepted that the fibre diameter (measured in microns) is the most dominant. Micron is the unit of length equivalent to one millionth of a metre or 0.001 mm. As the diameter determines the type of product the wool can be turned into, it is very often the first and arguably the most important characteristic an auction buyer would take into account.

Table 1.1: Descriptions of Significant Wool Characteristics

| Variables | Label | Description |
|---|---|---|
| Diameter | DIAMETER | The fibre diameter of wool. Micron is often used generically in the wool trade to describe the diameter: unit of length equivalent to one millionth of a metre or 0.001 mm. |
| Proportion of Breaks (middle) | POBMID | Measure of the percent of staples that broke in the middle third of the staple. Position of break is determined from the weight of the two ends of the broken staple. |
| Staple Length | SL | Measurement of the average length of wool staples. |
| Staple Strength | SS | Measure of the strength of a wool staple. Computed from peak force to break divided by linear density of the staple. |
| Vegetable Matter Content | VMB | Vegetable matter base is the dry weight of vegetable matter (burr, seed, hard heads etc.) expressed as a percentage of the weight of the greasy sample tested. |
| Yield | YIELD | As applied to greasy wool, yield is the estimate of the clean fibre either after washing/scouring or processing. |

Each bale of wool that goes through the auction system, whether sold or not, has its information recorded and archived by the Australian Wool Exchange (AWEx). Further, the sale lots that are sold also have their sale prices recorded. Table 1.2 below gives an idea of the data recorded. Once recorded, the prices paid in this auction market can potentially be used by the Australian production and service sectors to identify the quality preferences of the international retail markets and the intermediate processors. Auction price indicators and some premium and discount tables are released each week by the Australian Wool Exchange (AWEx) for the local market participants and for the international sectors that rely on their raw wool supplies from Australia. AWEx prepares market reports, which contain price tables. However, these tables are often sparse when several quality characteristics are combined.

Table 1.2: An Example of Data Recorded

| Sale Lot ID | DIAMETER (micron) | POBMID | SL | SS | VMB | YIELD | CPRICE (cents/kg) |
|---|---|---|---|---|---|---|---|
| 1 | 20.1 | 64 | 73 | 29 | 1.2 | 59.8 | 602.01 |
| 2 | 20.2 | 53 | 65 | 37 | 2.0 | 52.0 | 576.92 |
| 3 | 19.0 | 57 | 72 | 26 | 4.3 | 54.6 | 961.54 |
| 4 | 20.3 | 59 | 77 | 26 | 0.8 | 70.5 | 567.38 |
| 5 | 21.8 | 60 | 76 | 26 | 1.2 | 55.8 | 498.21 |
| 6 | 23.1 | 38 | 84 | 22 | 1.9 | 54.0 | 461.11 |
| 7 | 20.4 | 37 | 70 | 32 | 3.7 | 47.3 | 505.29 |
| 8 | 19.7 | 49 | 65 | 34 | 2.2 | 56.4 | 721.63 |
| 9 | 23.1 | 50 | 87 | 36 | 3.0 | 55.4 | 476.53 |
| 10 | 18.9 | 61 | 64 | 33 | 2.4 | 59.8 | 953.18 |
| . | | | | | | | |
| . | | | | | | | |
| . | | | | | | | |

The most commonly used tool for assessing the wool price and the effects of changes in quality on the wool price was “Pricemaker” (available from the Woolmark Company), and later Woolcheque. The data used in such software packages was supplied by AWEx.

The wool characteristics that were taken into account by the Pricemaker were: fibre diameter, staple strength, vegetable matter content, staple length, unscourable colour and style. The relative importance of these characteristics appeared to change over time, but there have been few attempts to identify the

relative importance or separate the effects of supply changes from changes in market demand. The only observation was that diameter dominates. Stanton (1993, 1994), Stanton and Coss (1995) and Stanton, et al. (1997) attempted to isolate these changes with the objective of achieving a fully informed auction market for Australian wool.

When using the Pricemaker application, the grower may choose to use those prices prevailing at Australian wool auctions during the previous selling season or during the most recent month.

Not much other work has been done on wool auction prices. There have been a number of studies involving agricultural forecasting (Allen 1994; Bessler 1994; Freebairn 1994; Tomek 1994), and very few involve the wool market and auctions (Graham-Higgs et al. 1999; Jones et al. 2004; Kemp and Willetts 1996; Simmons and Hansen 1997). In the past, Kemp and Willets (1996) have studied the remembering power of farmers or brokers in the Canterbury wool industry in New Zealand. They found that there is a general tendency to underestimate more recent and overestimate 14-year-old prices. On the other hand, Simmons and Hansen (1997) developed a theoretical model of the wool market that could distinguish between large and small buyers. The model forecasts that "large" buyers possess cost advantage, which will increase grower prices providing there is competition from periphery of small, relative high cost buyers. Graham-Higgs et al. (1999) found that the futures wool market is efficient for up to a six-month spread, but no further into the future. Because futures market prices can be used to predict prices up to six months in advance, wool growers can use the futures price to assess when they market their clip, but not for longer-term.

### 1.3 Problems Faced by the Wool Industry

The lack of clear market signals on the relative importance of wool attributes with respect to the price they receive at auction, has been an ongoing problem faced by wool growers. If the structure of Australian wool auction prices can be modelled adequately, then we can optimise the information that can be extracted

and used by the production and service sectors in the production and distribution of the raw wool clip. The more accurate modelling and analysis would help wool growers to better understand the market behaviour. If the important factors are identified, then effective strategies can be developed to maximise return to the growers.

Pricemaker (and later Woolcheque), like most wool price estimation systems being employed in the industry, is based on a series of multiple-linear and non-linear regressions. However, these regression systems have numerous assumptions and problems including non-linearity, non-normal distributions, interactions in causing price changes, correlations between characteristics and prediction resulted with negative price. The prediction from Pricemaker also have range limitations for given independent variables. In particular, prediction was limited within the range of 18.5 and 24.5 microns in fibre diameter. In addition, as the complexity of the regression models are increased, the problems of sparse data, and the need for aggregation of data over time to complete the datasets, became limiting.

There is also an industry assumption that all the measured characteristics (listed above) are relevant to all lots. Hence variables such as vegetable matter content are included as discounts in all sale lots, but at the same time the industry acknowledged that price discounts were relevant for only part of the population of sale lots.

The use of regression systems also have the assumption that the regression will pass through the average price, and that premiums and discounts are applied equally above and below the average. Prices from auctions do not present normally distributed observations above and below the average price. Elements of auction theory suggest that the distribution will be asymmetrical, or even a compound distribution, which has high frequencies of observations below the theoretical frontier price (the maximum expected price for a sale lot), and price tapering off as the technical inefficiencies accumulate within the sale lot.

So there is a need to explore the Australian wool price and its hedonic modelling relating changes in quality to changes in price. This needs to be undertaken with dual objectives of addressing the information needs of the industry participants, and addressing and overcoming the limitations of the current regression systems.

Additionally, there also exists a wool specifications problem. Industrial sorting of wool during harvest, and at the start of processing, assembles wool in bins according to the required wool specifications. At present this assembly is done by constraining the range of all specifications in each bin, and having either a very large number of bins, or a large variance of characteristics within each bin. Multiple linear regression on price does not provide additional useful information that would streamline this process, nor does it assist in delineating the specifications of individual bins.

## 1.4 Summary of Our Research, and Overview of This Thesis

In this thesis, we address the problems faced by the Australian wool growers as presented in Section 1.3. We will present our search for alternative means to model the structure of Australian wool auction prices, as well as an approach that would provide useful information in streamlining the wool specifications process as detailed at the end of Section 1.3.

The more accurate modelling and analysis helps wool growers to better understand the market behaviour. If the important factors are identified, then effective strategies can be developed to maximise return to the growers.

In Chapter 1, we presented a brief overview of the Australian wool auction market. We discussed the problems faced by the wool growers and their significance, which motivated our research.

In Chapter 2, we define the predictive aspect of the modelling problem and present the data that is available to us for our research. We introduce the

assumptions that must be made in order to model the auction data and predict the wool prices.

Chapter 3 discusses neural networks and their potential in our wool auction problem. Neural networks are known to give good results in many modern applications resolving industrial problems (Cheng and Titterington, 1994, Frost and Karri, 1999). We perform an analysis and assessment of neural networks, specifically, the generalised regression neural networks (GRNN). We look at their strengths and weaknesses, and apply GRNN to the wool auction problem and comment on their relevance and usability in our wool problem. We find that Neural Networks such as GRNN work well with our wool auction data and give excellent prediction results with good accuracies. However, they do not provide any clear picture to the understanding of the interactions between the various price-driving wool variables. As such, we continue to investigate alternative methods which would provide a clear picture and better interpretability of the wool variables.

Chapter 4 details the tree-based regression methods, as an alternate approach to neural networks. We find that a regression tree provides a very clear picture of the relative importance of each price-driving wool variable at various levels, in the form of a single tree diagram. Further, we develop an alternative tabular representation of the tree diagram, which solves the wool specification first discussed in Section 1.3, and is of immerse use to wool growers. However, we find that the prediction accuracies of regression trees are relatively poor compared to GRNN. We investigate the ensemble methods of improving the prediction accuracies of trees such as bagging and random forests. While such methods do improve the prediction accuracies when applied on our wool auction data, they bring complexity to the interpretability of the wool variables by introducing multiple tree diagrams. We can certainly combine and average the numerical outputs from multiple tree diagrams, but we cannot combine and take average of the tree diagrams themselves. The simplicity and clarity of working with only a single tree is lost from the ensemble methods. Hence, we need to develop a method which provides prediction accuracies comparable to those of GRNN, but at the same time retain the same level of interpretability of a single

tree. Although the ensemble methods are not ideal, they lead to an excellent idea for our work in Chapter 5.

Chapter 5 details the new hybrid approach we developed as a result of the work we did in Chapter 3 and 4 using neural networks and tree-based regression method. Our hybrid approach combines the two methods with their respective strengths. We apply our new approach to the data, compare the results with our earlier work in neural networks and tree-based regression methods, and discuss the results. We find that our new hybrid approach is the best balance between prediction accuracies and interpretability that we can currently achieve. It provides solutions to the various aspects of our initial wool auction problem, as well as the wool specifications problem discussed in Section 1.3.

Finally, we conclude our thesis with Chapter 6, discussing the potential of our new hybrid approach and the directions of possible future works.

# Chapter 2

# Data and Assumptions

In this chapter we will define the predictive aspect of the modelling problem and present the data that is available to us for our research. We will introduce the assumptions that must be made in order to model the auction data and predict the wool prices.

## 2.1 Prediction

There are two aspects of modelling: predictive and descriptive. They are both required to be satisfied for our wool auction price problem. The predictive (and fitting) aspect is quantitative and can be directly compared across different models with relative ease. The descriptive aspect, however, is qualitative and cannot be compared numerically across different models. We will comment on the descriptive aspect of different models individually in their respective chapters. In the following, we first define the typical problem of prediction.

A predictive problem (or regression) is easy to state but difficult to solve in general. Say a quantifiable attribute $y$ (called the "output" or "response" variable) is assumed to be dependent on a vector of attributes/characteristics/properties $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ (often called "input" variables). In other words, there exists an underlying but unknown function $f$ between $\boldsymbol{x}$ and $y$:

$$y = f(x_1, x_2, \ldots, x_n) = f(\boldsymbol{x}) \tag{2.1}$$

The goal then is to find, construct, or fit a predictor function $\hat{f}$ that most closely approximates $f$ from a training set of size $N$: $\{(\boldsymbol{x}(n), y(n)) : n = 1, \ldots, N\}$, a finite

set of possibly noisy measured/observed values of $\boldsymbol{x}$ and the associated values of $y$. With $\hat{f}$ we can compute/predict (estimate) $\hat{y}$ (the most probable value of $y$) for each value of $\boldsymbol{x}$ :

$$y \sim \hat{y} = \hat{f}(\boldsymbol{x}) \tag{2.2}$$

or

$$y = \hat{f}(\boldsymbol{x}) + (\text{noise and error}) \tag{2.3}$$

The problem then is to find a "good" predictor or "best fit" such that the overall error is minimised. Often the root mean square error can be used to represent this overall error.

Our problem differs from typical problems of a statistics nature. In statistics, most often a small sample is used for reaching some form of conclusion about a much larger population that is too large to be worked on. Our problem is closer to data mining in nature, where we attempt to model all data (if available) in a single period and use the conclusion to predict outcomes for the next period; i.e. we build a model with the current week's wool auction data to predict the prices in next week's auction.

## 2.2 Assumptions

Since the buyers in a wool auction make their bidding/purchasing decisions based on the laboratory measured wool quality characteristics released to them, it makes sense for us to use these as the input variables in our model to predict the price. There exists a large set of wool characteristics, and we shall consider a subset that is most widely accepted in the industry as the most significant, namely: the fibre diameter, proportion of break (middle), staple length, staple strength, vegetable matter content, and yield. We make our first assumption that the price of wool ($y$) can be expressed as an unknown function $f$, such that:

$$y = f(x_1, x_2, \ldots, x_6) = f(\boldsymbol{x}) \tag{2.4}$$

where

$x_1$ = DIAMETER

$x_2$ = POBMID

$x_3$ = SL

$x_4$ = SS

$x_5$ = VMB

$x_6$ = YIELD

Now, knowing the auctions are held frequently and almost weekly, one would expect the auctions that are held close together (say within a few weeks) would have their price structures/patterns affected by mostly the same factors, more so than auctions that are further apart (say two months or more) which would be much more different due to market changes over the longer period of time. One would expect the difference in price structure/pattern between two consecutive auctions to be relatively small and almost negligible. Hence, if we are to predict the prices of wool in a particular auction, we can assume it would be sufficient to build a predictive model from data from a selection of past auctions that are reasonably close to the one being predicted.

We shall apply the above assumptions for all models within this thesis.

Of course, auction prices, like share and oil prices, depend not only on the product specifications and historical behaviour but also on intangible factors such as speculations, international market influences, and unexpected social and political events. To make our predictions as accurate as possible, ideally the intangible factors should be identified and captured in our models, and their influences analysed. For our research, we make the decision to focus on building models from the limited data that is readily available to us. Considerations of the intangible factors will certainly strengthen our models and should be considered in future research.

## 2.3 Data Considered

Our research partner, the Department of Agriculture and Food, Government of Western Australia, has kept an extensive data collection of wool that were offered in auctions from the late 1960s to the present. Figure 2.1 shows the average price per kilogram of clean wool from Fremantle auction data between July 1998 and December 2002. Series 1 represents wool having a fibre diameter of 19 microns, Series 2 represents wool having a fibre diameter of 21 microns, while Series 3 represents wool having a fibre diameter of 23 microns.

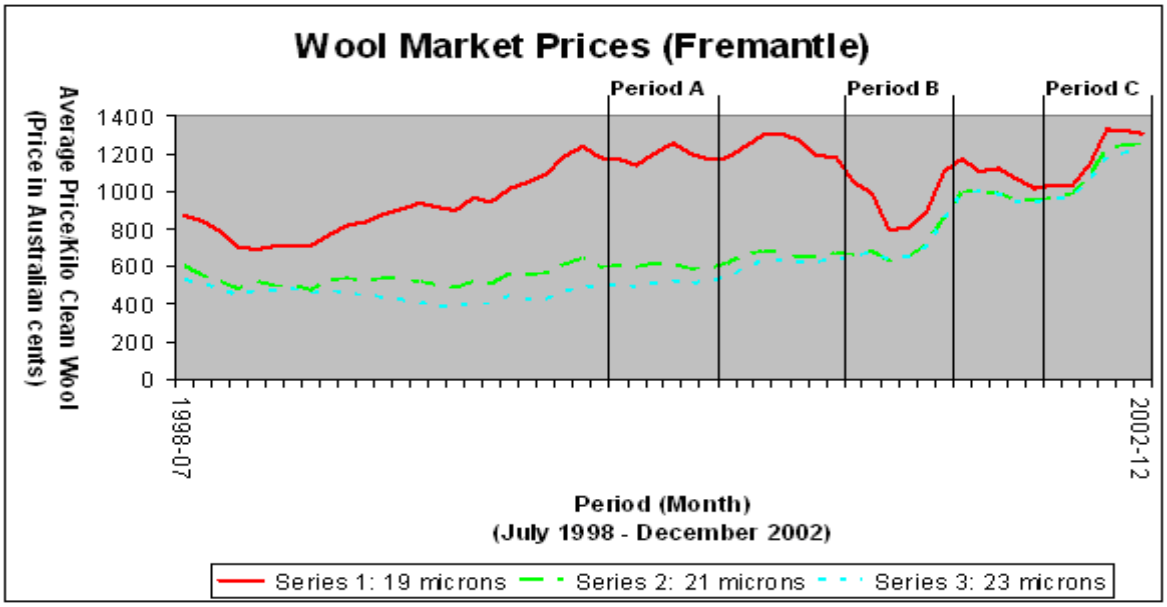


Figure 2.1: Three Periods of Interest

As mentioned in Section 1.2, the fibre diameter (measured in microns) is well regarded by the industry as the most important price-driving wool characteristic. This is reflected in the majority of Figure 2.1, where finer wool was sold for a higher average price. However, it can be seen that this price difference starts to diminish sometime around August 2001.

For our research, the Department of Agriculture and Food provided us data of Merino fleece wool from Fremantle auctions in the three separate periods as shown in Figure 2.1: July 2000 – December 2000 (Period A), August 2001 – January 2002 (Period B), and July 2002 – December 2002 (Period C). Period A

is of interest to the department because the price difference due to diameter was the greatest. Period B is interesting because this was when the price difference began to diminish. And finally in Period C the price difference due to diameter was minimal.

As mentioned in Section 1.1, auctions are held almost every week in the year, but with breaks during the Easter period, a period in July, and the Christmas period. Depending on the volume of wool sale lots in each auction, the auction can be spread over 1 to 3 days. To avoid the breaks and to have some form of uniformity when comparing the three periods, we shall only consider the months August to December in each of the three periods. The details are given in Table 2.1.

Table 2.1: Weekly Auctions Held in the Three Periods

| | Period A (2000) | Period B (2001) | Period C (2002) |
|---|---|---|---|
| August | 5 Days in 3 Weeks: | 6 Days in 3 Weeks: | 6 Days in 3 Weeks: |
| Week 1 | 16/8 ,17/8 | 14/8, 15/8, 16/8 | 14/8, 15/8 |
| Week 2 | 23/8 | 22/8, 23/8 | 21/8, 22/8 |
| Week 3 | 30/8, 31/8 | 29/8 | 28/8, 29/8 |
| September | 8 Days in 4 Weeks: | 5 Days in 4 Weeks: | 8 Days in 4 Weeks: |
| Week 1 | 6/9, 7/9 | 5/9 | 4/9, 5/9 |
| Week 2 | 13/9, 14/9 | 13/9 | 11/9, 12/9 |
| Week 3 | 20/9, 21/9 | 19/9 | 18/9, 19/9 |
| Week 4 | 27/9, 28/9 | 26/9, 27/9 | 25/9, 26/9 |
| October | 7 Days in 4 Weeks: | 8 Days in 4 Weeks: | 9 Days 4 Weeks: |
| Week 1 | 5/10 | 9/10, 10/10, 11/10 | 8/10, 9/10, 10/10 |
| Week 2 | 11/10, 12/10 | 17/10, 18/10 | 16/10, 17/10 |
| Week 3 | 18/10, 19/10 | 24/10, 25/10 | 23/10, 24/10 |
| Week 4 | 25/10, 26/10 | 31/10 | 30/10, 31/10 |
| November | 8 Days in 4 Weeks: | 5 Days in 3 Weeks: | 6 Days in 3 Weeks: |
| Week 1 | 1/11, 2/11 | 7/11, 8/11 | 6/11, 7/11 |
| Week 2 | 8/11, 9/11 | 21/11, 22/11 | 20/11, 21/11 |
| Week 3 | 21/11, 22/11, 23/11 | 28/11 | 27/11, 28/11 |
| Week 4 | 29/11 | | |
| December | 3 Days in 2 Weeks: | 3 Days in 2 Weeks: | 4 Days in 2 Weeks: |
| Week 1 | 6/12 | 5/12 | 4/12, 5/12 |
| Week 2 | 13/12, 14/12 | 12/12, 13/12 | 11/12, 12/12 |

## 2.4 Model Comparisons

As illustrated in Table 2.1, the number of auctions varies in a month across the three periods. To maintain some form of uniformity, we shall only model the last week worth of auction data from each month, then use the model to predict the price outcomes in the first week of the next month.

Table 2.2: Number of Sale Lots in Our Data

| | Period A (2000) | Period B (2001) | Period C (2002) |
|---|---|---|---|
| Last week of August | 1503 | 1236 | 1344 |
| 1st week of September | 1847 | 1406 | 1321 |
| Last week of September | 1890 | 2216 | 1461 |
| 1st week of October | 1302 | 2632 | 2895 |
| Last week of October | 1806 | 1050 | 1293 |
| 1st week of November | 2157 | 1581 | 1607 |
| Last week of November | 1160 | 922 | 903 |
| 1st week of December | 1042 | 863 | 1252 |

To compare how "good" or how accurate a model is against another, we will use the typical measures of:

- root mean square error,
- mean absolute error, and
- standard deviation of absolute error

in the rest of this thesis.

In the next chapter, we will discuss neural networks and specifically, the generalised regression neural networks (GRNN), which has been found to give satisfactory predictions of wool auction prices by our research partner, the Department of Agriculture and Food, WA. We will look at the strengths and weaknesses of neural networks, and use the results from GRNN as the benchmark in our comparisons of different modelling methods in the rest of this thesis.

# Chapter 3

# Neural Networks

In this chapter, we will discuss artificial neural networks as alternate methods in modelling and predicting the Australian wool auction prices. Neural networks are known to give good results in many modern applications resolving industrial problems (Cheng and Titterington, 1994, Frost and Karri, 1999). As a result of the popularity of such methods and the ongoing development of them, our research partner, the Department of Agriculture and Food, Government of Western Australia, performed a preliminary investigation into neural networks and found them to give satisfactory predictions of wool auction prices.

In the following sections, we will perform an analysis and assessment of neural networks, specifically, the generalised regression neural networks (GRNN). We will look at the strengths and weaknesses of GRNN, and apply GRNN to the wool auction problem and comment on their relevance and usability in our wool problem. We will detail the problems we face, and why neural networks may not be a good choice for the wool auction problem. We will also use the numerical prediction results from GRNN as the benchmark in our comparisons of different modelling methods in the rest of this thesis.

This chapter concludes that Neural Networks such as GRNN work well with our wool auction data and give excellent prediction results with good accuracies. However, they do not provide any clear picture to the understanding of the interactions between the various price-driving wool variables. As such, we continue to investigate alternative methods which would provide a clear picture and better interpretability of the wool variables. We will use the prediction results from GRNN as a benchmark when comparing the prediction accuracies across various methods, and investigate the feasibility of integrating GRNN into some of these methods, if such possibility arises.

## 3.1 Why Neural Networks?

The relationships between the price and price-driving wool attributes are non-linear, interactive and very complex. Neural networks can readily handle non-linearity with ease and without restrictions, and thus are much more attractive than earlier multiple-linear models mentioned in Section 1.3. Also, there exist other issues such as the price and wool attribute relationships being dynamic over time, and the wool data set available in a given period could be incomplete and imprecise. The flexibilities and ongoing developments of neural networks allow researchers to explore and make continual advancements in the handling of such issues.

## 3.2 Neural Networks Basics

There is no universally accepted definition of a neural network. However, most researchers concede that a neural network is a network composed of a large number of simple processors (neurons) that are massively interconnected, operate in parallel, and learn from experience (examples). These are the primary known characteristics of biological neural systems that are the easiest to exploit in artificial neural systems.

The inspiration for neural nets comes from the structure of the brain. A brain consists of a large number of cells, referred to as "neurons". A neuron receives impulses from other neurons through a number of "dendrites". Depending on the impulses received, a neuron may send a signal to other neurons, through its single "axon", which connects to dendrites of other neurons. Like the brain, the structure of an artificial neural net consists of connected units referred to as "nodes" or "neurons". Each neuron performs a portion of the computations inside the net: a neuron takes some numbers as inputs, performs a relatively simple computation on these inputs, and returns an output. The output value of a neuron is passed on as one of the inputs for another neuron, except for neurons that generate the final output values of the entire system.

Neurons are arranged in layers. The input layer neurons receive the inputs for the computations, like the diameter, proportion of breaks (middle), staple length, staple strength, vegetable matter content, and yield of an individual wool sale lot. These values are passed to the neurons in the first hidden layer, which perform computations on their inputs and pass their outputs to the next layer. This next layer could be another hidden layer, if there is one. The outputs from the neurons in the last hidden layer are passed to the neuron or neurons that generate the final outputs of the net, like the auction price of the wool sale lot.

Training a net is the process of fine-tuning the parameters (weights) of the computation, where the purpose is to make the net output approximately correct values for the given inputs. This process is guided by training data on the one hand, and the training algorithm on the other. The training algorithm selects various sets of computation parameters, and evaluates each set by applying the net to each training case to determine how good the answers given by the net are. Each set of parameters is a "trial"; the training algorithm selects new sets of parameters based on the results of previous trials. In other words, neural networks "learn" from examples, as children learn to distinguish dogs from cats based on examples of dogs and cats. If trained carefully, neural networks may exhibit some capability for generalisation beyond the training data, that is, to produce approximately correct results for new cases that were not used for training.

We have investigated the use of neural networks with the Department of Agriculture and Food, WA in predicting wool prices for the industry. In particular, GRNN (Generalized Regression Neural Networks) has been considered because of its availability and speed.

GRNN (Specht 1991) is a memory-based network that provides estimates of continuous variables and converges to the underlying (linear or nonlinear) regression surface. This type of neural network is a one-pass learning algorithm with a highly parallel structure. Even with sparse data in a multidimensional measurement space, the algorithm provides smooth transitions from one

observed value to another. The algorithm form can be used for any regression problem in which an assumption of linearity is not justified. The parallel network form has found use in applications such as learning the dynamics of a plant model for prediction or control.

With GRNN there is no need for the user to make decisions about the structure of a net. These nets always have two hidden layers of neurons, with one neuron per training case in the first hidden layer (the Pattern Layer), and two neurons in the second layer (the Summation Layer). A GRNN for p independent numeric variables is structured as shown in the graph displayed in Figure 3.1.

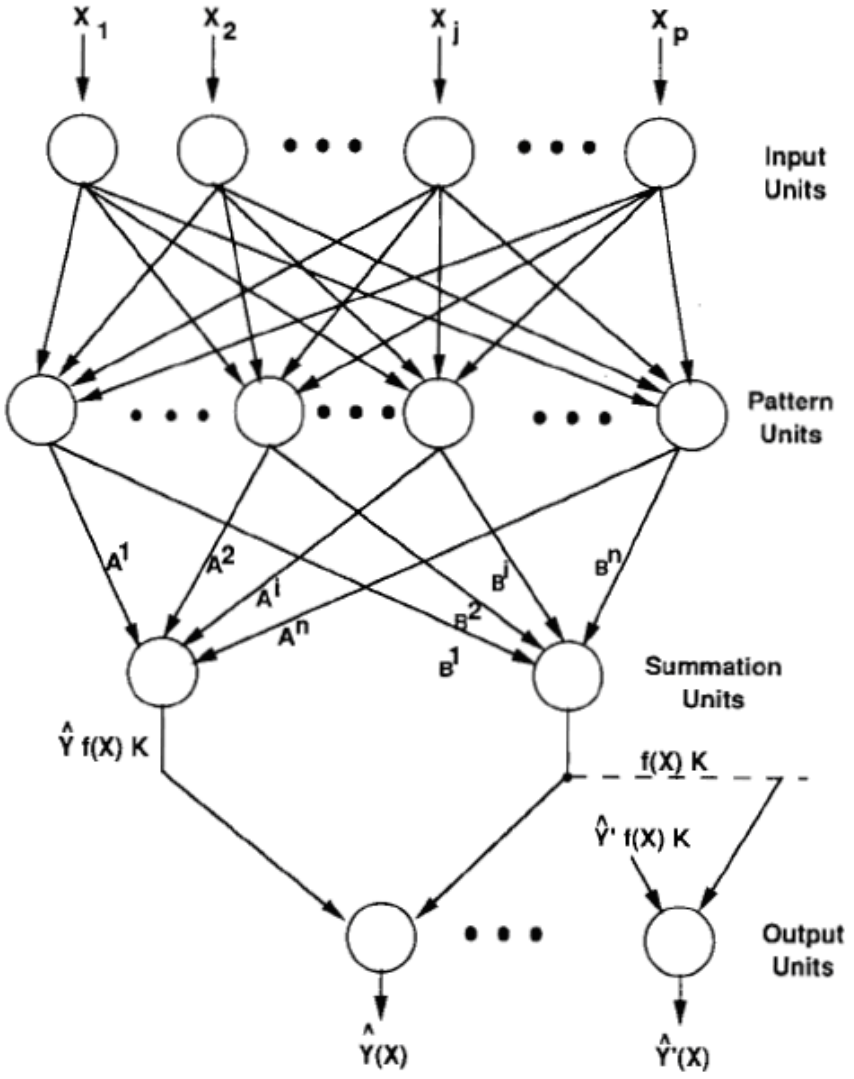


Figure 3.1: GRNN for p Independent Numeric Variables (Specht 1991)

The Pattern Layer contains one node for each training case. Presenting a training case to the net consists here of presenting p independent numeric values. Each neuron in the pattern layer computes its distance from the presented case. The values passed to the two nodes in the Summation Layer (Numerator and Denominator Nodes) are functions of the distance and the dependent value. The Numerator and Denominator Nodes in the Summation Layer sum its inputs, while the Output Node divides them to generate the prediction.

The distance function computed in the Pattern Layer neurons uses "smoothing factors"; every input has its own "smoothing factor" value. With a single input, the greater the value of the smoothing factor, the more significant distant training cases become for the predicted value. With 2 inputs, the smoothing factor relates to the distance along one axis on a plane, and in general, with multiple inputs, to one dimension in multi-dimensional space.

Training a GRNN consists of optimizing smoothing factors to minimize the error on the training set, and the Conjugate Gradient Descent optimization method is used to accomplish that. The error measure used during training to evaluate different sets of smoothing factors is the Mean Squared Error. However, when computing the Squared Error for a training case, that case is temporarily excluded from the Pattern Layer. This is because the excluded neuron would compute a zero distance, making other neurons insignificant in the computation of the prediction.

Advantages of GRNN include:

- Train fast.
- Does not require topology specification (numbers of hidden layers and nodes).

Disadvantages of GRNN include:

- Bigger in size than some neural nets such as MLF (Multi-Layer Feed-forward), thus slower to make predictions.

- Less reliable outside the range of training data (for example, when the value of some independent variables falls outside the range of values for that variable in the training data); though note that prediction outside the range of training data is still risky with other neural nets.
- Lower capability of generalising from very small training sets.

GRNN is a universal approximator for smooth functions, so it should be able to solve any smooth function-approximation problem given enough data.

### 3.3 Applying Neural Networks to the Wool Auction Data

In this section we provide an assessment of the neural networks' ability to model auction data and predict wool prices. We consider GRNN in particular because of its availability and speed. In our entire thesis, all GRNN neural nets used are generated from the commercially available software package from Palisade called NeuralTools. This is the same software packaged used by the Department of Agriculture and Food. NeuralTools is available as an Excel add-on package under the Windows environment. However, the version of Excel used in this project (Excel 2003) has a limitation of allowing 65536 rows of data in a file. While this does not affect our work in the current chapter, we will comment on this limitation in other chapters of our thesis.

The network topology follows the standard GRNN as shown in Figure 3.1. As a standard package, NeuralTools follows Specht (1991) closely in determining both the network topology and the weights of the links in the background. By default, 80% of a data set is used for training and the remaining 20% is used for testing. The tolerance for bad predictions is set at 30% for both training and testing.

In the following, we assess GRNN's ability to model (or fit) the data, as well as the accuracy in prediction. We use the data introduced in Section 2.3, with the three periods described by Figure 2.1: August to December 2000 (Period A), August to December 2001 (Period B), and August to December 2002 (Period C),

covering the dates listed in Table 2.1. We model the last week worth of auction data from each month, then use the model to predict the price outcomes in the first week of the following month. For example, we first generate a GRNN neural net to model (fit) the data of the last week in August 2000. Then we use this net to predict the auction price outcomes from the data of the first week of September 2000. We repeat this for all months available to us across the three periods. Our results follow.

Tables 3.1, 3.2 and 3.3 show the results from modelling (fitting) the last week of each month in Periods A, B and C with GRNN. In the same tables we also show the results generated from typical multiple linear regression (labelled MLR) for comparison. The Department of Agriculture and Wool, WA found our accuracies of fitting with GRNN to be very acceptable. Figures 3.2 to 3.13 on the following pages show the fitted price vs. actual price in the weeks we selected. The more accurate a fitting is, the more each plot appears as a straight line. It can be observed that the fittings in the three periods are very good in general with GRNN.

Table 3.1: Fitting for Period A with GRNN

| | | | MLR | GRNN |
|---|---|---|---|---|
| Period A | Fitting last week of Aug 2000 | Root Mean Square Error | 101.9485 | 28.29091354 |
| | | Mean Absolute Error | 71.97585 | 15.26685036 |
| | | Std. Deviation of Abs. Error | 72.22491 | 23.82596829 |
| | Fitting last week of Sep 2000 | Root Mean Square Error | 132.8114 | 33.47677404 |
| | | Mean Absolute Error | 97.1946 | 16.33235456 |
| | | Std. Deviation of Abs. Error | 90.53411 | 29.23013271 |
| | Fitting last week of Oct 2000 | Root Mean Square Error | 167.2928 | 35.59787003 |
| | | Mean Absolute Error | 121.3715 | 20.77791974 |
| | | Std. Deviation of Abs. Error | 115.1659 | 28.9127874 |
| | Fitting last week of Nov 2000 | Root Mean Square Error | 140.6487 | 53.14404304 |
| | | Mean Absolute Error | 105.0392 | 17.53775253 |
| | | Std. Deviation of Abs. Error | 93.57544 | 50.1885246 |

Table 3.2: Fitting for Period B with GRNN

| | | | MLR | GRNN |
|---|---|---|---|---|
| Period B | Fitting last week of Aug 2001 | Root Mean Square Error | 97.37026 | 35.52785964 |
| | | Mean Absolute Error | 73.29277 | 19.05397627 |
| | | Std. Deviation of Abs. Error | 64.1285 | 29.99838122 |
| | Fitting last week of Sep 2001 | Root Mean Square Error | 89.59667 | 43.38224682 |
| | | Mean Absolute Error | 60.11147 | 18.91764899 |
| | | Std. Deviation of Abs. Error | 66.45425 | 39.04907165 |
| | Fitting last week of Oct 2001 | Root Mean Square Error | 55.50737 | 29.53830859 |
| | | Mean Absolute Error | 39.87708 | 16.2291678 |
| | | Std. Deviation of Abs. Error | 38.63041 | 24.6922348 |
| | Fitting last week of Nov 2001 | Root Mean Square Error | 55.29734 | 21.18334939 |
| | | Mean Absolute Error | 36.27066 | 11.16225224 |
| | | Std. Deviation of Abs. Error | 41.76275 | 18.01361591 |

Table 3.3: Fitting for Period C with GRNN

| | | | MLR | GRNN |
|---|---|---|---|---|
| Period C | Fitting last week of Aug 2002 | Root Mean Square Error | 53.13123 | 27.4876211 |
| | | Mean Absolute Error | 34.05608 | 15.47762742 |
| | | Std. Deviation of Abs. Error | 40.79644 | 22.72436108 |
| | Fitting last week of Sep 2002 | Root Mean Square Error | 74.73919 | 39.05447769 |
| | | Mean Absolute Error | 50.48061 | 25.27669893 |
| | | Std. Deviation of Abs. Error | 55.13379 | 29.78166889 |
| | Fitting last week of Oct 2002 | Root Mean Square Error | 50.63217 | 33.21652489 |
| | | Mean Absolute Error | 36.25862 | 19.23416188 |
| | | Std. Deviation of Abs. Error | 35.35387 | 27.09155177 |
| | Fitting last week of Nov 2002 | Root Mean Square Error | 42.94687 | 33.74128347 |
| | | Mean Absolute Error | 29.83724 | 21.64586617 |
| | | Std. Deviation of Abs. Error | 30.90681 | 25.89736289 |

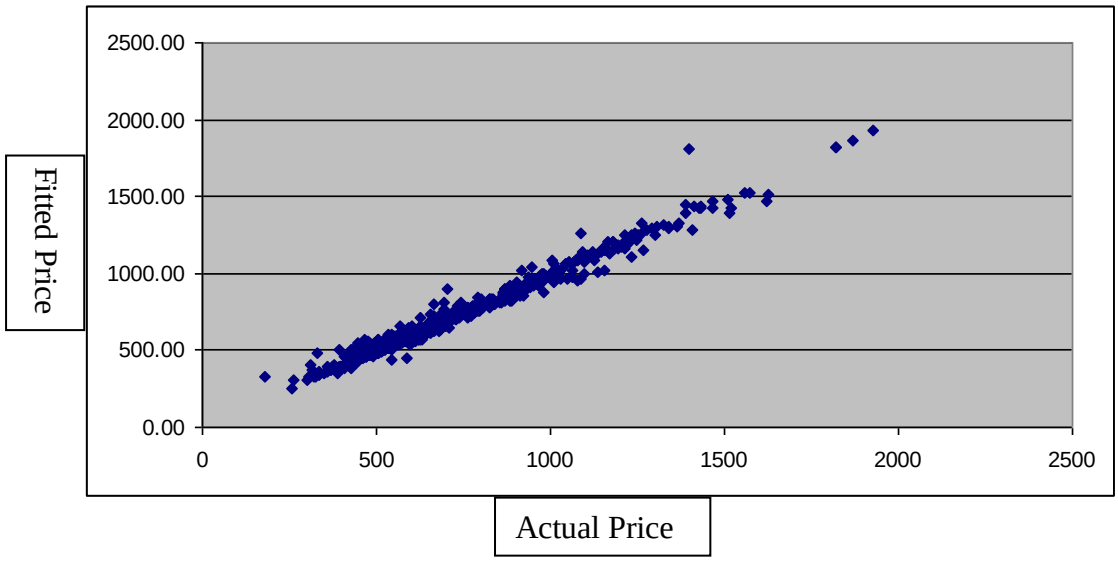


Figure 3.2: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of August 2000 with GRNN

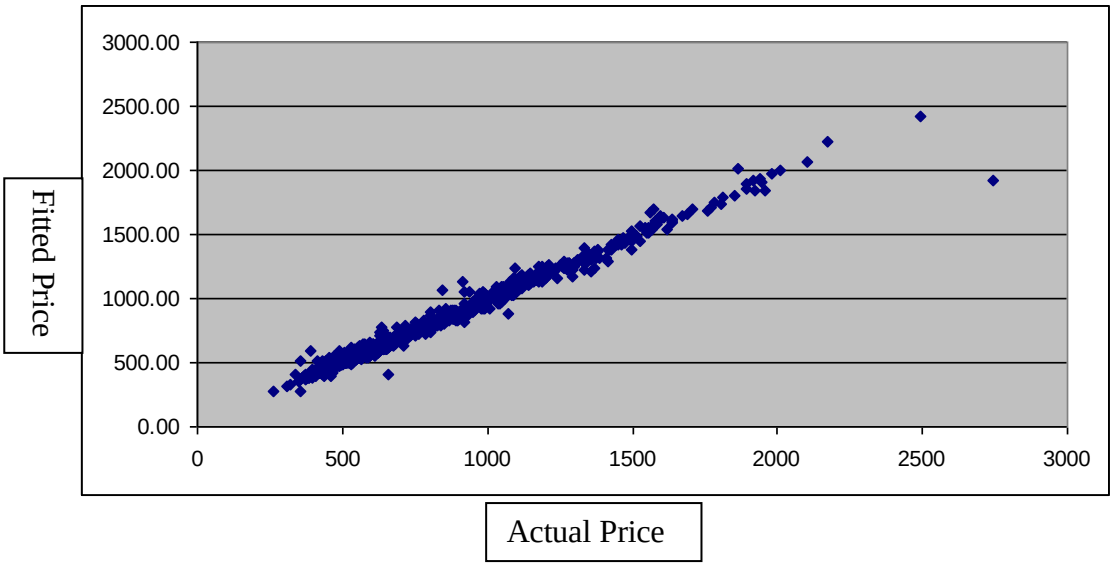


Figure 3.3: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of September 2000 with GRNN

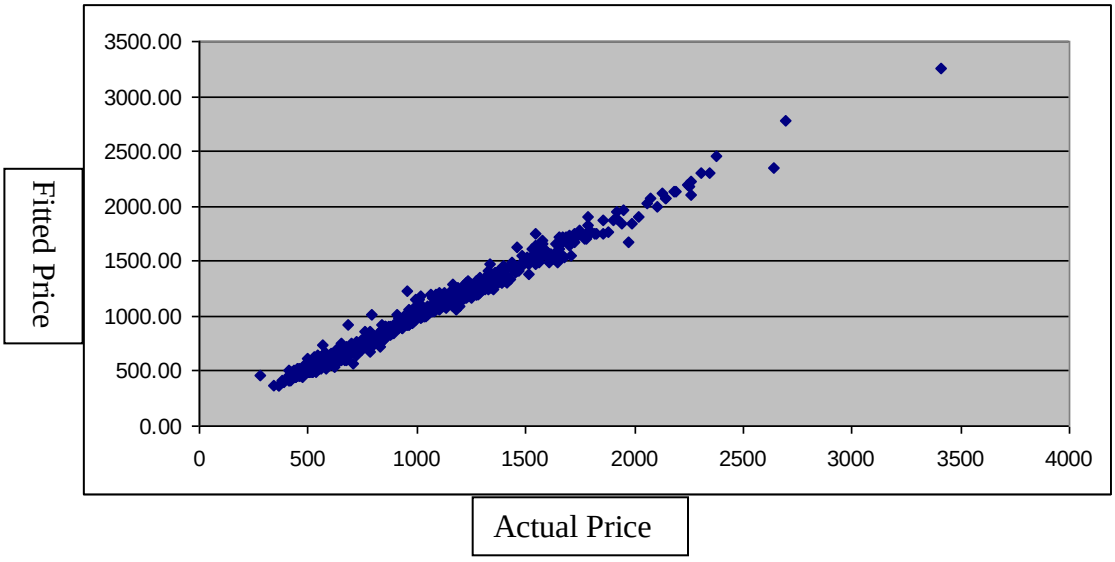


Figure 3.4: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of October 2000 with GRNN

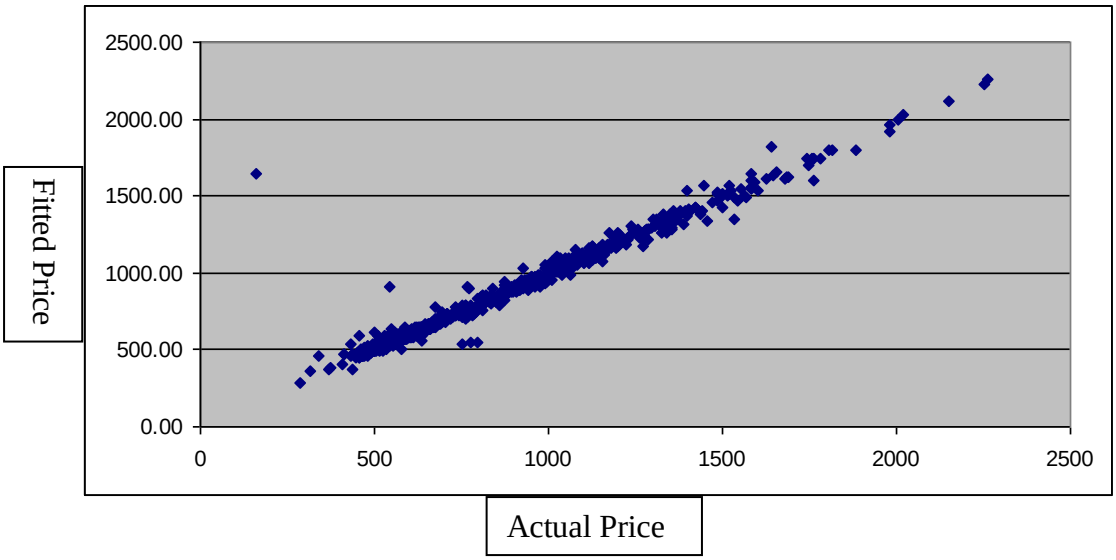


Figure 3.5: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of November 2000 with GRNN

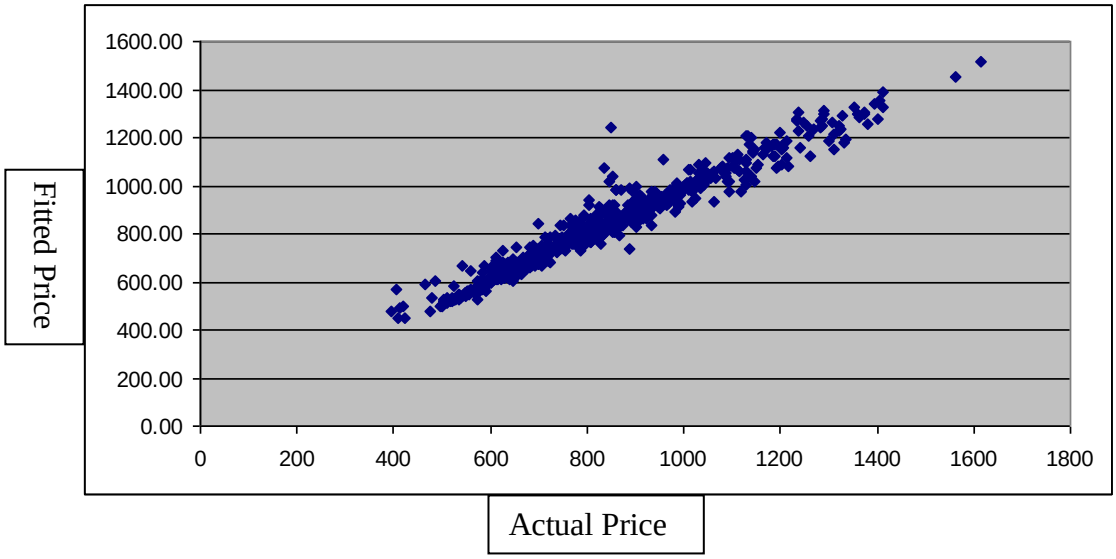


Figure 3.6: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of August 2001 with GRNN

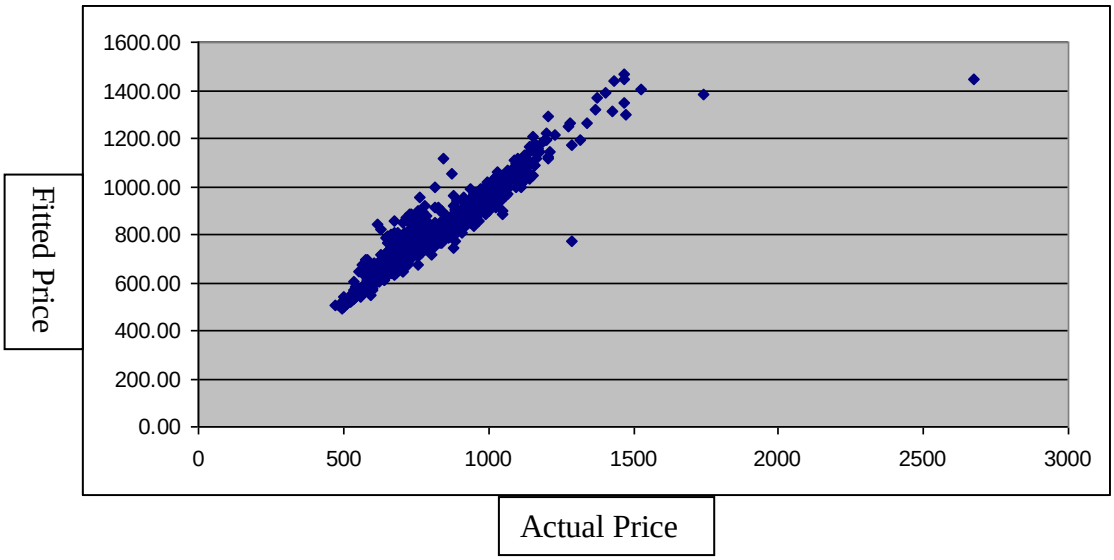


Figure 3.7: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of September 2001 with GRNN

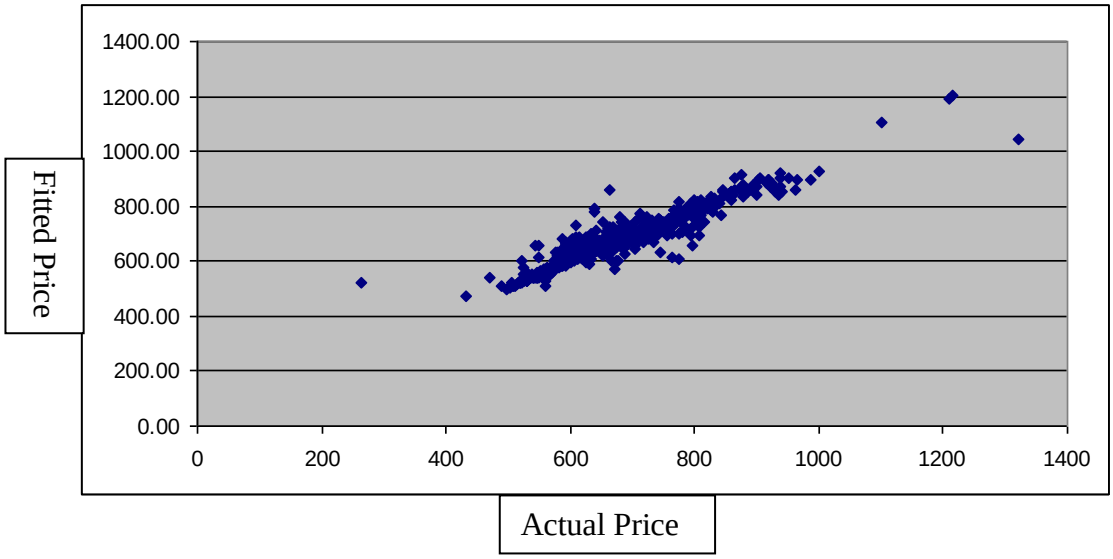


Figure 3.8: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of October 2001 with GRNN

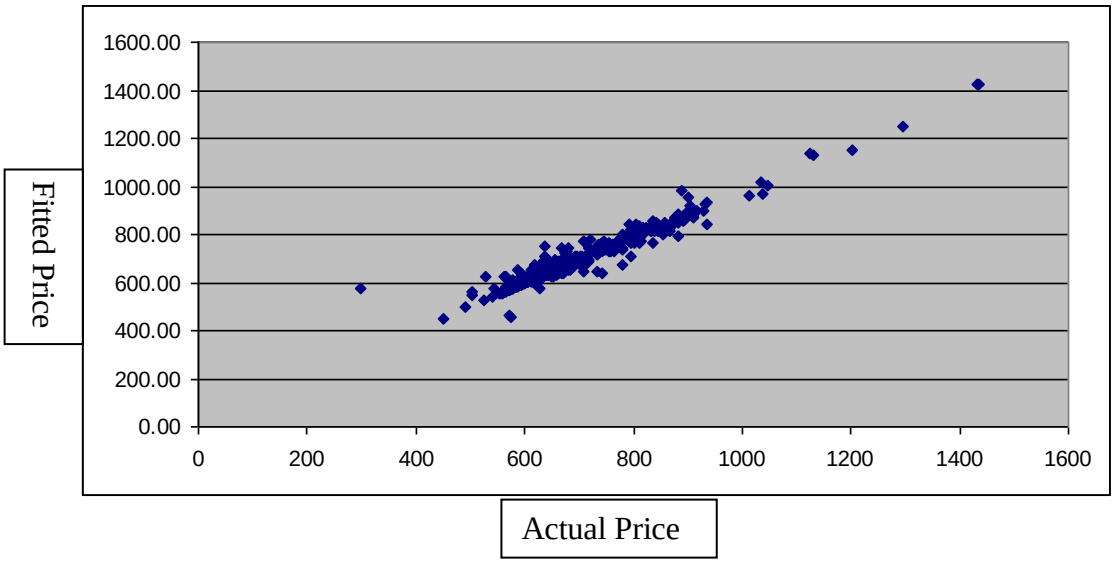


Figure 3.9: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of November 2001 with GRNN

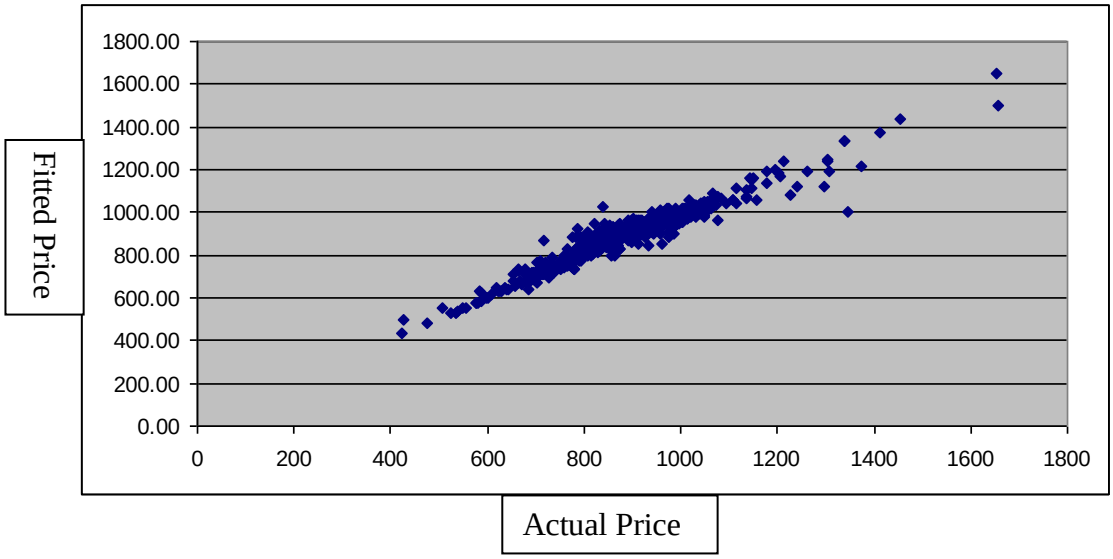


Figure 3.10: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of August 2002 with GRNN

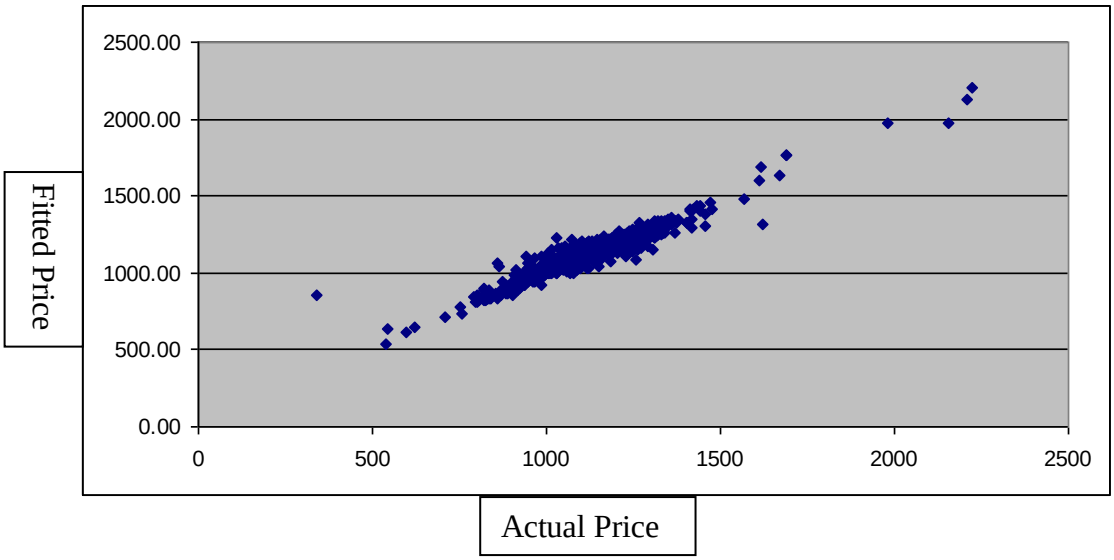


Figure 3.11: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of September 2002 with GRNN

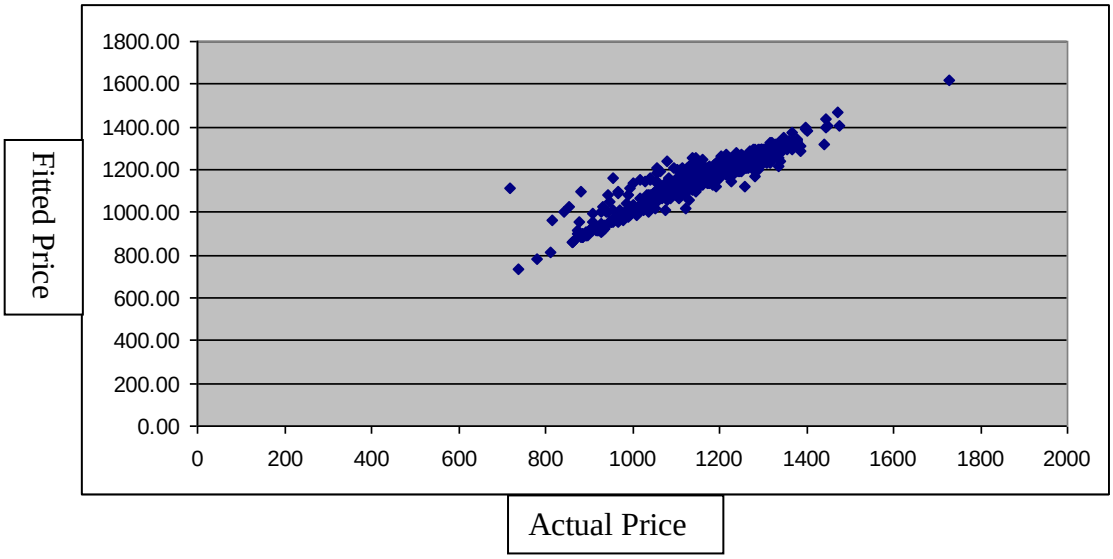


Figure 3.12: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of October 2002 with GRNN

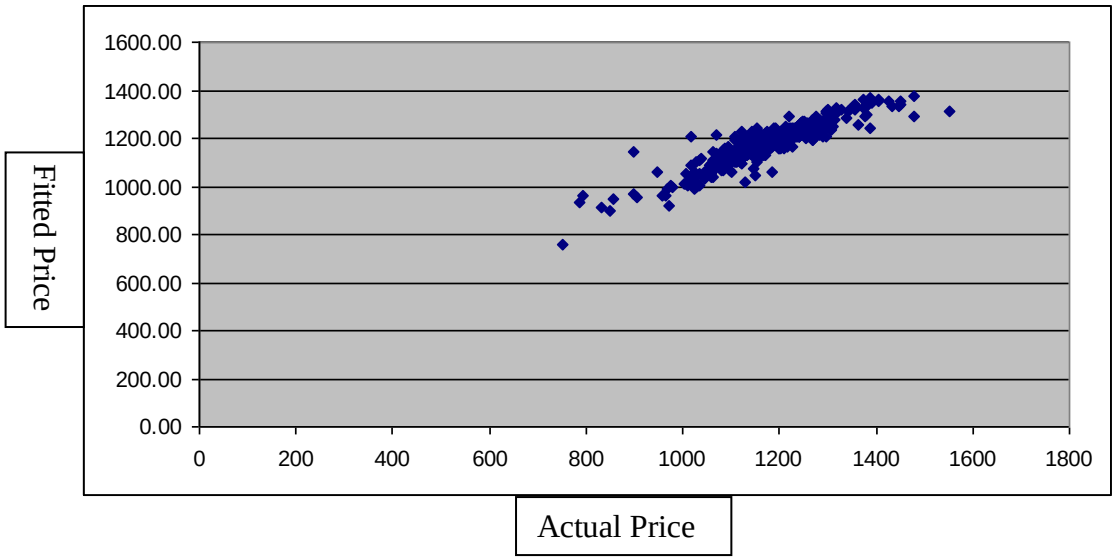


Figure 3.13: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of November 2002 with GRNN

After fitting the last week of a month with GRNN, we can now look at using the fitted model to predict the price outcomes in the first week of the following month. Tables 3.4, 3.5 and 3.6 on the following pages show the results from predicting the last week of each month in Periods A, B and C with the fitted model of the last week of previous month. We have found the errors of these to be acceptable for predictions. And Figures 3.14 to 3.25 that follow show the predicted price vs. actual price in the weeks we selected. It can be observed that the plots all follow straight lines and the predictions can be considered to be reasonably accurate.

Table 3.4: Predictions for Period A with GRNN

| | | | MLR | GRNN |
|---|---|---|---|---|
| Period A | Using last week of Aug 2000 to predict 1st wk of Sep 2000 | Root Mean Square Error | 124.9364 | 47.88411498 |
| | | Mean Absolute Error | 83.78764 | 27.49475198 |
| | | Std. Deviation of Abs. Error | 92.70047 | 39.21427868 |
| | Using last week of Sep 2000 to predict 1st wk of Oct 2000 | Root Mean Square Error | 209.4786 | 135.7182532 |
| | | Mean Absolute Error | 115.9219 | 29.63332056 |
| | | Std. Deviation of Abs. Error | 174.5473 | 132.4945037 |
| | Using last week of Oct 2000 to predict 1st wk of Nov 2000 | Root Mean Square Error | 188.5797 | 69.12056444 |
| | | Mean Absolute Error | 141.316 | 35.6166384 |
| | | Std. Deviation of Abs. Error | 124.8972 | 59.2514565 |
| | Using last week of Nov 2000 to predict 1st wk of Dec 2000 | Root Mean Square Error | 203.728 | 64.7849604 |
| | | Mean Absolute Error | 118.1895 | 34.43145383 |
| | | Std. Deviation of Abs. Error | 166.0204 | 54.90408942 |

Table 3.5: Predictions for Period B with GRNN

| | | | MLR | GRNN |
|---|---|---|---|---|
| Period B | Using last week of Aug 2001 to predict 1st wk of Sep 2001 | Root Mean Square Error | 110.5063 | 58.66584341 |
| | | Mean Absolute Error | 83.0728 | 32.43893849 |
| | | Std. Deviation of Abs. Error | 72.89942 | 48.8988455 |
| | Using last week of Sep 2001 to predict 1st wk of Oct 2001 | Root Mean Square Error | 105.0846 | 89.2013982 |
| | | Mean Absolute Error | 86.1434 | 74.55923056 |
| | | Std. Deviation of Abs. Error | 60.19528 | 48.97674904 |
| | Using last week of Oct 2001 to predict 1st wk of Nov 2001 | Root Mean Square Error | 57.24911 | 38.35475621 |
| | | Mean Absolute Error | 42.80339 | 27.41881763 |
| | | Std. Deviation of Abs. Error | 38.02954 | 26.82817574 |
| | Using last week of Nov 2001 to predict 1st wk of Dec 2001 | Root Mean Square Error | 84.60831 | 66.14964581 |
| | | Mean Absolute Error | 60.50812 | 54.72642547 |
| | | Std. Deviation of Abs. Error | 59.17255 | 37.18058425 |

Table 3.6: Predictions for Period C with GRNN

| | | | MLR | GRNN |
|---|---|---|---|---|
| Period C | Using last week of Aug 2002 to predict 1st wk of Sep 2002 | Root Mean Square Error | 47.72842 | 38.60712508 |
| | | Mean Absolute Error | 34.57202 | 27.09162225 |
| | | Std. Deviation of Abs. Error | 32.91804 | 27.51594552 |
| | Using last week of Sep 2002 to predict 1st wk of Oct 2002 | Root Mean Square Error | 77.06622 | 67.49464254 |
| | | Mean Absolute Error | 52.55659 | 42.78633291 |
| | | Std. Deviation of Abs. Error | 56.37468 | 52.20917584 |
| | Using last week of Oct 2002 to predict 1st wk of Nov 2002 | Root Mean Square Error | 60.4721 | 53.65067084 |
| | | Mean Absolute Error | 44.47539 | 38.96359524 |
| | | Std. Deviation of Abs. Error | 40.98609 | 36.89281363 |
| | Using last week of Nov 2002 to predict 1st wk of Dec 2002 | Root Mean Square Error | 53.04425 | 41.68142591 |
| | | Mean Absolute Error | 36.6779 | 28.36965042 |
| | | Std. Deviation of Abs. Error | 38.33533 | 30.54913432 |

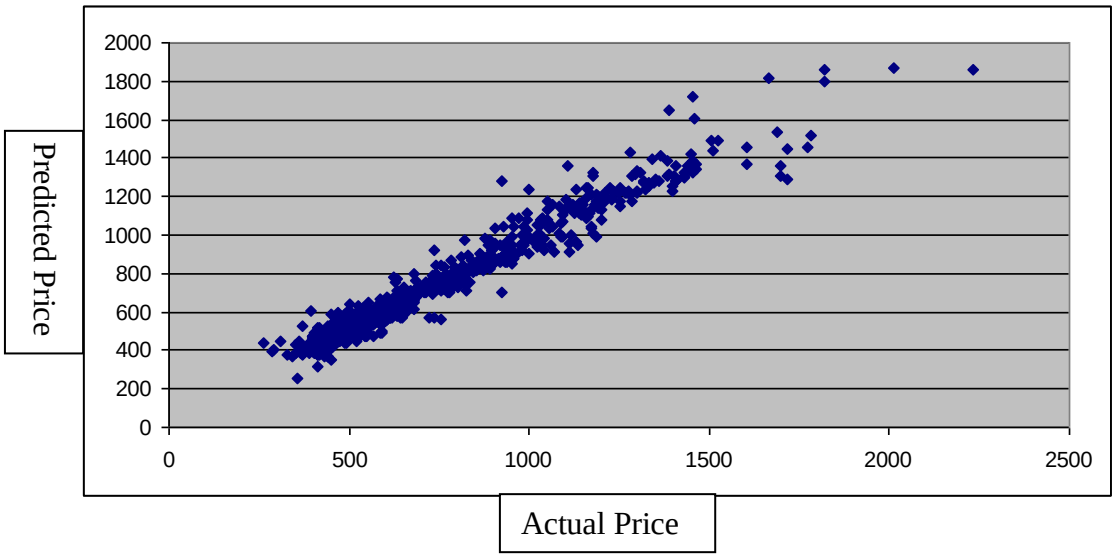


Figure 3.14: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the first week of September 2000 with GRNN

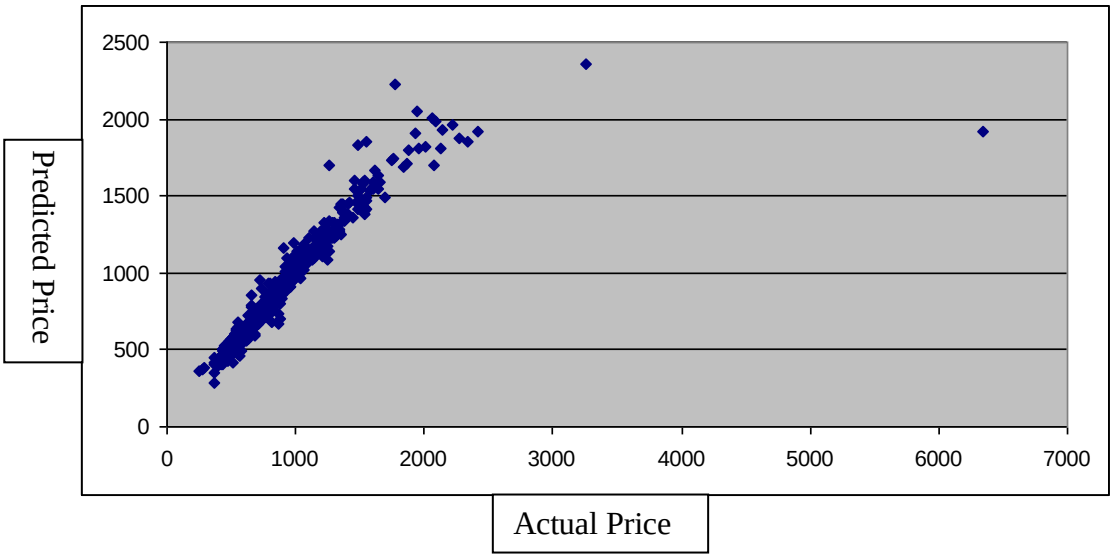


Figure 3.15: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the first week of October 2000 with GRNN

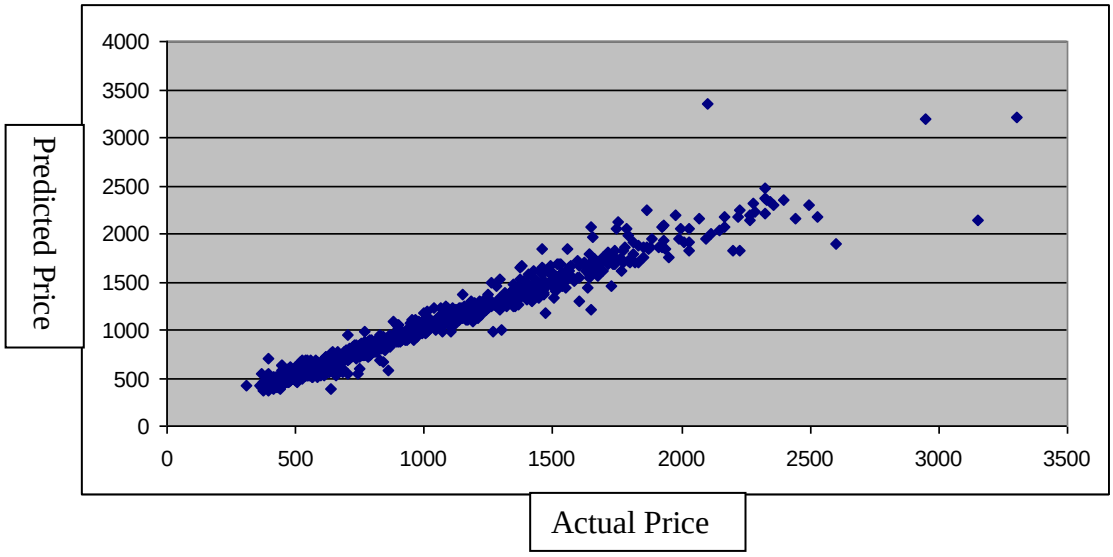


Figure 3.16: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the first week of November 2000 with GRNN

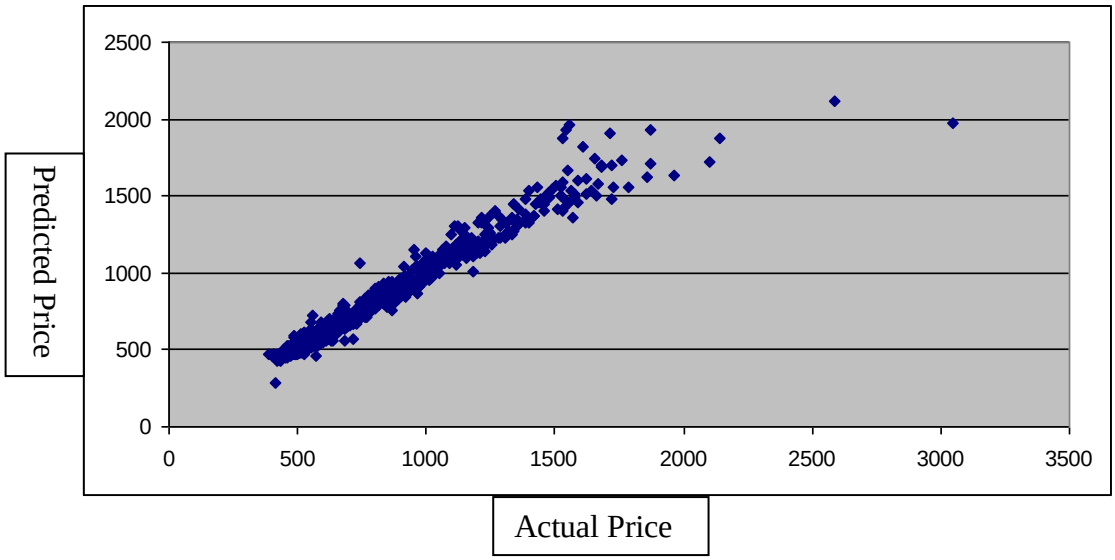


Figure 3.17: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the first week of December 2000 with GRNN

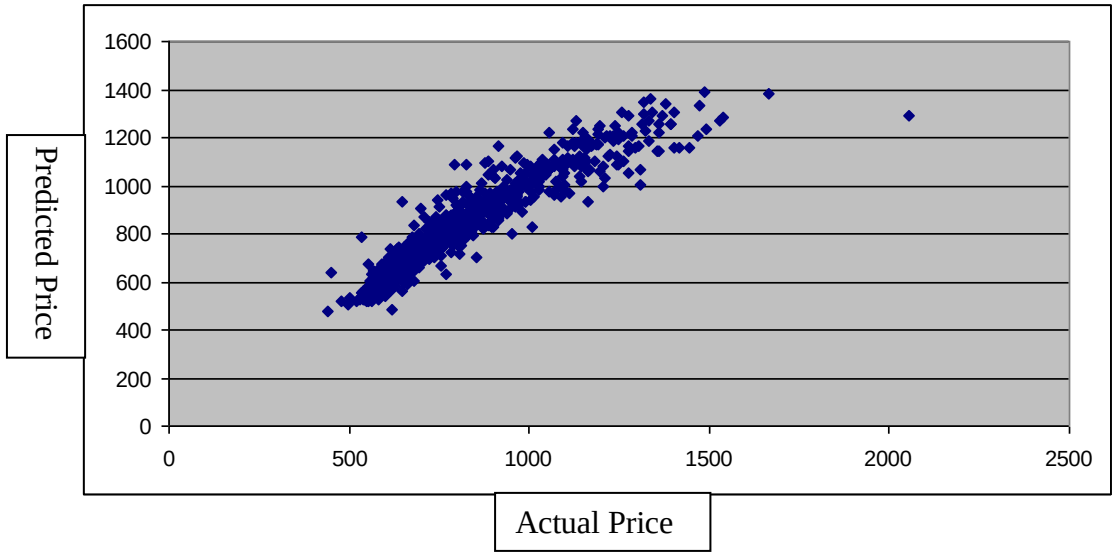


Figure 3.18: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the first week of September 2001 with GRNN

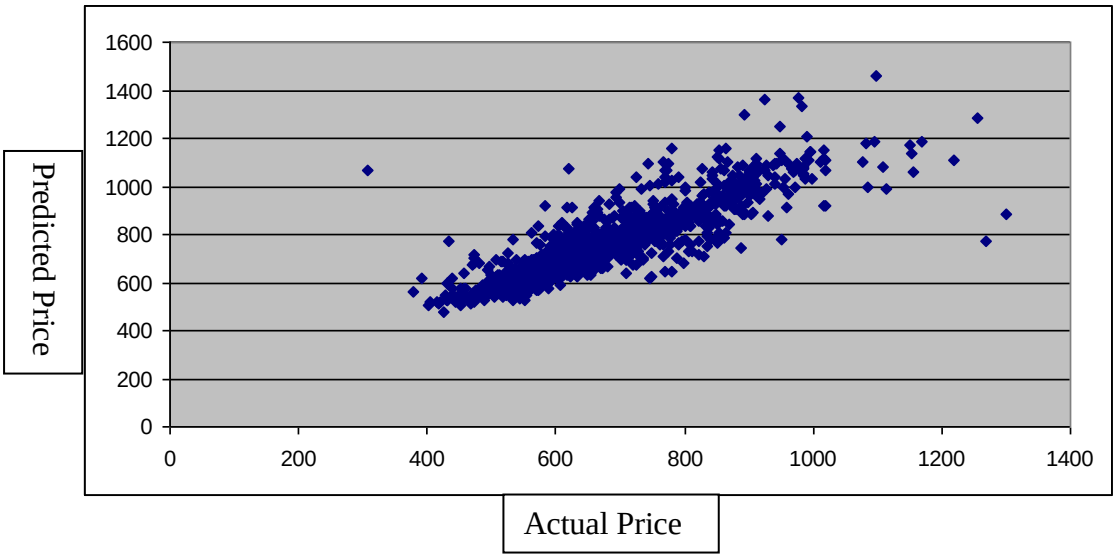


Figure 3.19: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the first week of October 2001 with GRNN

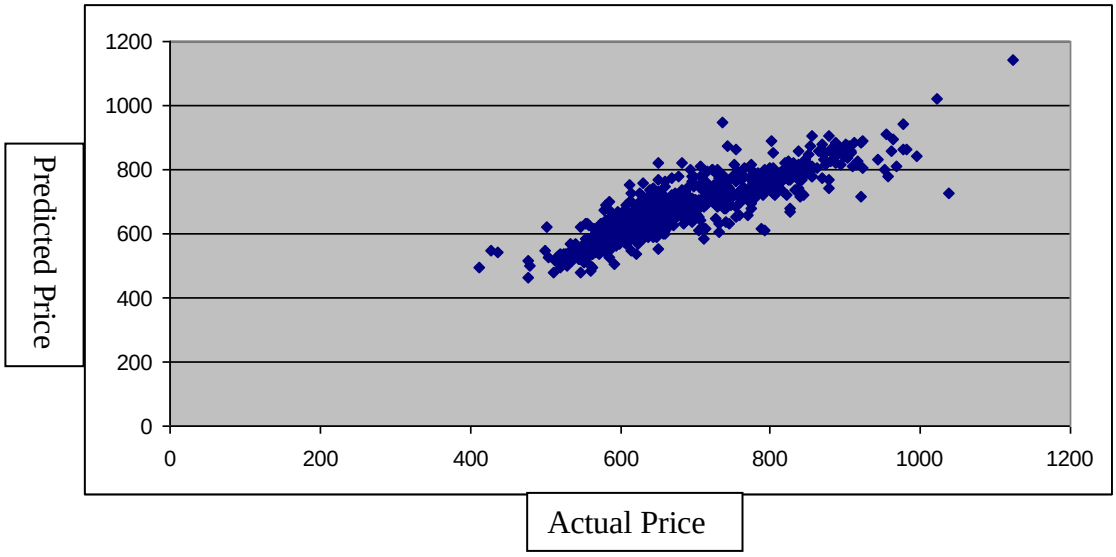


Figure 3.20: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the first week of November 2001 with GRNN

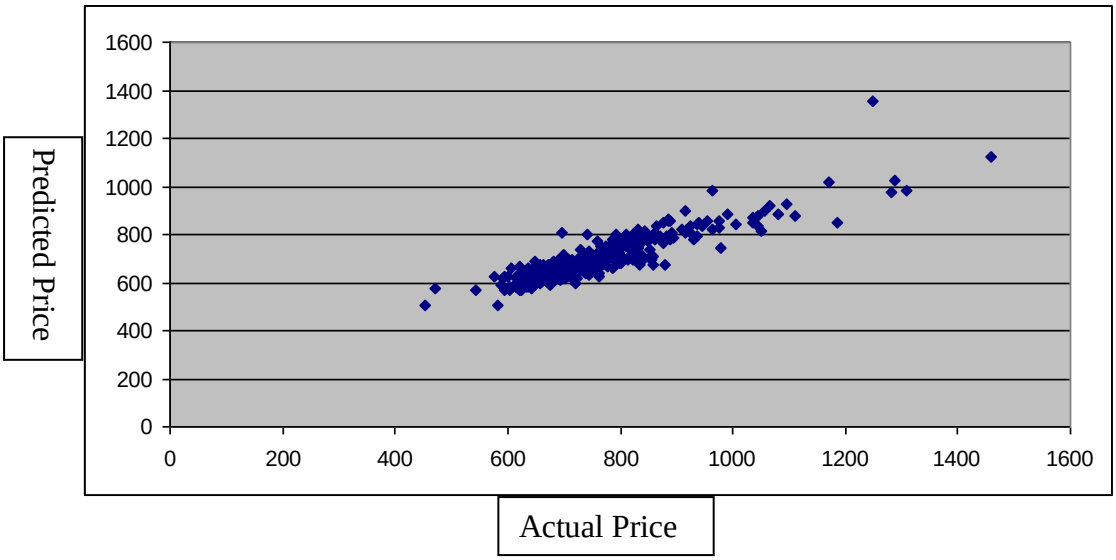


Figure 3.21: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the first week of December 2001 with GRNN

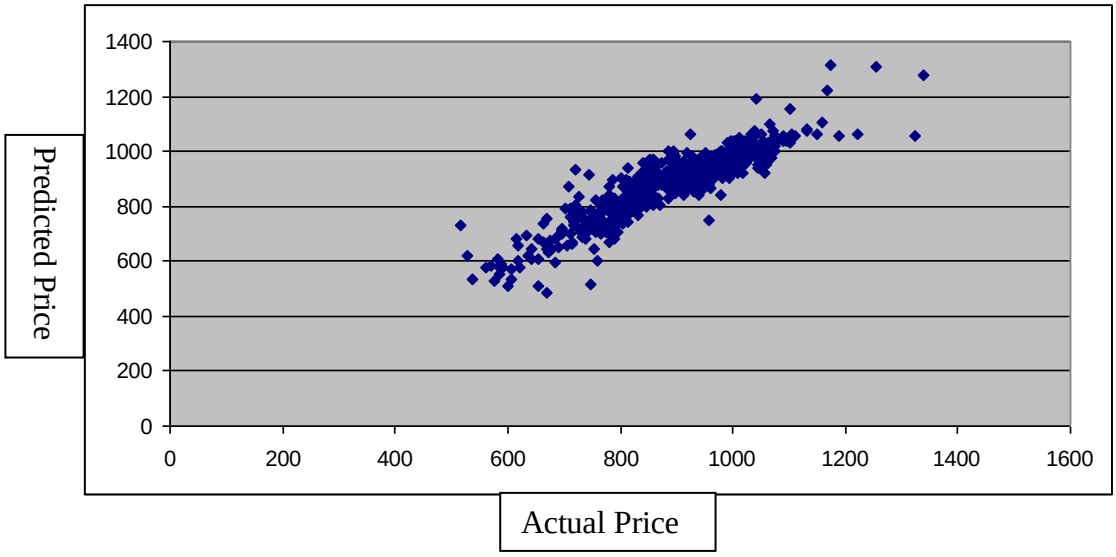


Figure 3.22: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the first week of September 2002 with GRNN

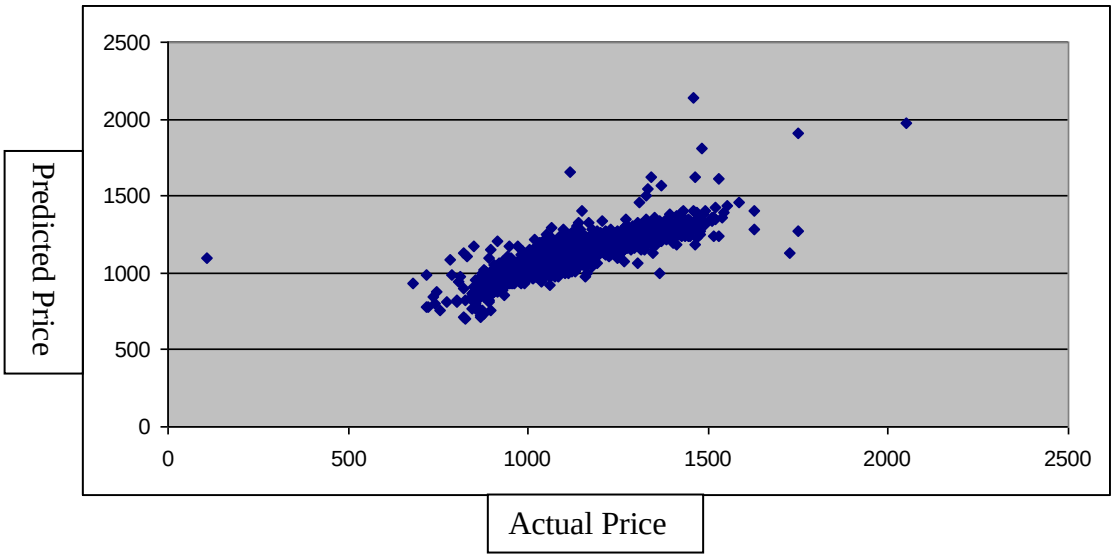


Figure 3.23: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the first week of October 2002 with GRNN

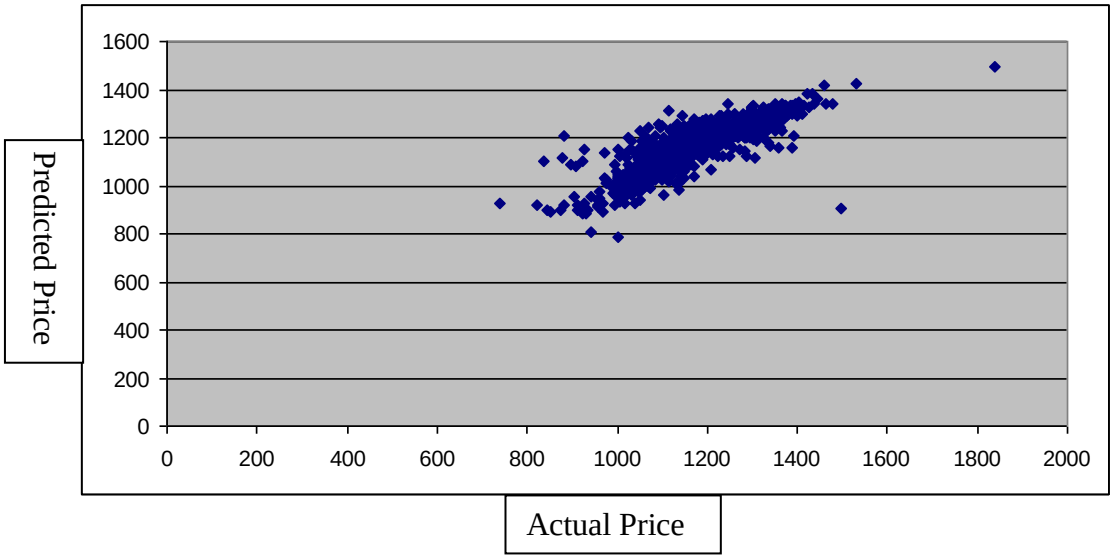


Figure 3.24: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the first week of November 2002 with GRNN

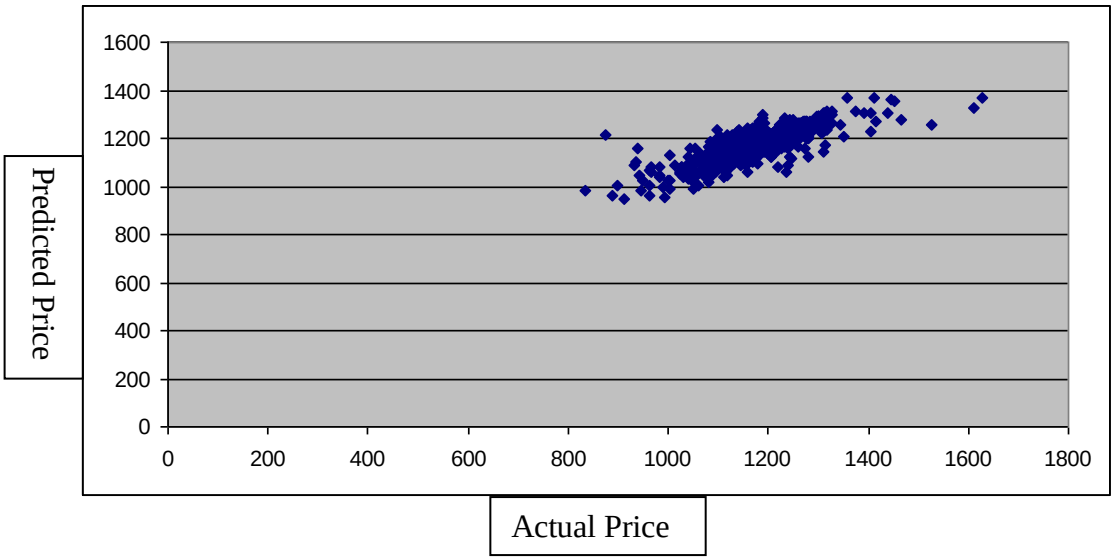


Figure 3.25: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the first week of December 2002 with GRNN

## 3.4 Discussions on Neural Networks

From the results in Section 3.3 we find the plots of fitted/predicted price vs. actual price to follow straight lines, thus we find the modelling (fitting) capacity and prediction accuracies of neural networks (GRNN) to be excellent. Neural networks satisfy the fitting/predictive aspect of modelling, and we shall use the results from Section 3.3 as our "benchmarks" in comparing the goodness of fit and prediction accuracies with other modelling methods in the rest of this thesis. However, being a "black box" method, neural networks can provide no further interpretation or actual understanding of the price-driving variables. Because of this, we need to investigate alternative methods that would satisfy the descriptive aspect of modelling and give us a better interpretation of the interaction between the wool quality characteristics and final price. We will also investigate the feasibility of integrating GRNN into some of these methods, if such possibility arises. In the next chapter, we will present our investigation of tree-based regression methods and explore their abilities to model the wool auction data.

# Chapter 4

# Tree-based Regression Methods

In this chapter, we will detail the tree-based regression methods, first with simple regression trees. We will illustrate their advantages over neural networks, as well as the trade-offs, in modelling the wool auction data. We will then consider the ensemble methods such as bootstrap aggregating (bagging) and random forests, and discuss their results.

This chapter concludes that a regression tree provides a very clear picture of the relative importance of each price-driving wool variable at various levels, in the form of a single tree diagram. Further, within this chapter we develop an alternative tabular representation of the tree diagram, which solves the wool specification first discussed in Section 1.3. However, we find that the prediction accuracies of regression trees are relatively poor compared to GRNN. We investigate the ensemble methods of improving the prediction accuracies of trees such as bagging and random forests. While such methods do improve the prediction accuracies when applied on our wool auction data, they bring complexity to the interpretability of the wool variables by introducing multiple tree diagrams. We can certainly combine and average the numerical outputs from multiple tree diagrams, but we cannot combine and take average of the tree diagrams themselves. The simplicity and clarity of working with only a single tree is lost from the ensemble methods. Hence, we need to develop a method which provides prediction accuracies comparable to those of GRNN, but at the same time retain the same level of interpretability of a single tree.

## 4.1 Regression Tree and Its Advantages

Trees have a universal simplicity. It is appealing to try and find tree representations of more complex relationships in a difficult problem. Cheng et

al. (2002) proposed to model the Australian wool auction prices by using tree-based regression. They initially modelled data between July 2000 and December 2000 in Fremantle using regression trees and compared the results with those from older industry methods. They found that regression trees had the clear advantage of being able to detect interaction between parts of levels or parts of the numeric range of independent variables. Figure 4.1 below gives an example of a regression tree.

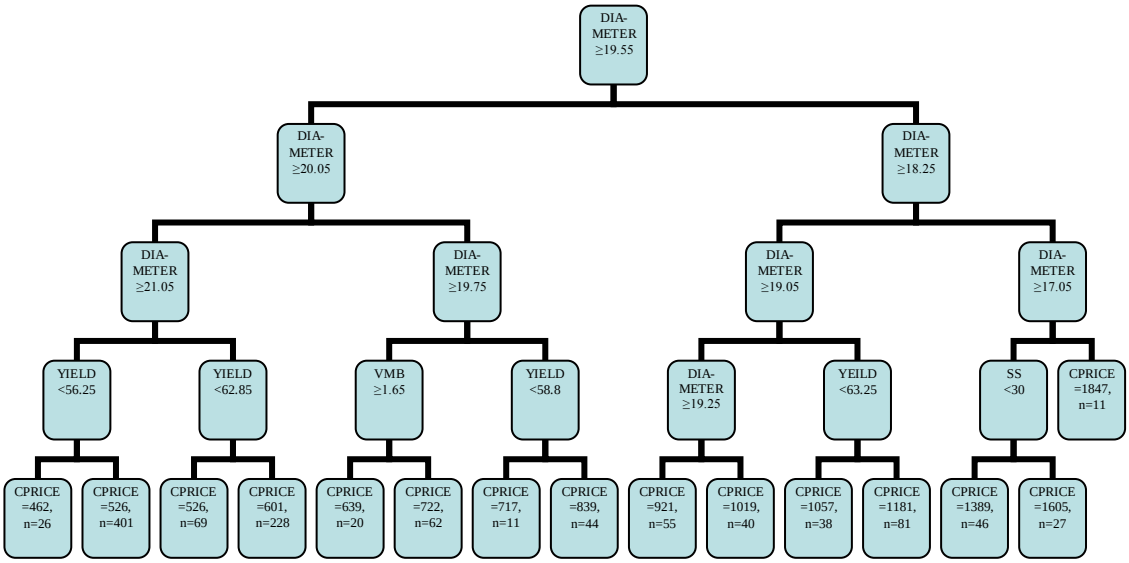


Figure 4.1: Example of a Regression Tree

The biggest advantage of a regression tree over neural networks is the tree diagram that is exclusive to the tree method. A tree diagram can be interpreted and give us better understanding of the price-driving variables, satisfying the descriptive aspect of modelling. The diagram can show us the order of importance of the wool quality variables and their influences in driving the price at various levels.

The automatic construction of a decision tree was first used in the social sciences field by Morgan and Sonquist (1963). Tree-based method (Breiman et al., 1984) is an alternative means to generalised linear (Watters and Deriso, 2000) and additive models for regression problems and to linear logistic and additive logistic models for classification problems. These types of models are fitted by

binary recursive partitioning of a dataset into increasingly homogeneous subsets until it is infeasible to continue. Their use in other fields such as social science (Morgan and Sonquist, 1963, Morgan and Messenger, 1973), statistics (Breiman et al., 1984) and machine learning (Quinlan, 1979, 1983 and 1986) has been widespread.

Tree-based models are defined by the algorithm used to fit them and belong to statistical classification techniques. The algorithm partitions the space of independent variables (**X**) into homogeneous regions such that, within each region, the conditional distribution of $y$ given $x$, $f(y|x)$, does not depend on $x$. Independent variables can be of several types: factors and numeric.

The deviance of a regression tree is the usual scaled deviance for a linear model, namely:

$$D = \sum_j (y_j - \mu_j)^2 , \qquad (4.1)$$

where $y_j \sim N(0, \sigma^2)$, $i = 1, \ldots, N$, is the response and $\mu_j = \rho(x_j)$. The split that gives the largest reduction in deviance will be chosen.

For reasons having to do with ease of computations, this "average squared error" is the measure of accuracy classically used in regression. The methodology revolving about this measure is the least squares (LS) regression. Alternatively, Breiman et al. (1984) also considered the use of least absolute deviation (LAD) regression, and they found LAD regression to work well with certain data sets.

The tree-based regression is fitted using binary recursive partitioning whereby the data are successively split along coordinate axes of the predictor variables so that at any node, the split that maximally distinguishes the response variable in the left and the right branches is selected. Splitting continues until nodes are pure or data are too sparse; terminal nodes are called leaves, while the initial node is called the root. If the response variable is numeric, the tree is called a regression tree. The model used for regression assumes that the numeric response variable has a normal (Gaussian) distribution.

There are many advantages to tree-based models over classical linear models. Tree-based regressions are easier to interpret and discuss in contrast to linear models when analysing a set of independent variables that contain a mixture of numeric variables and factors. In addition, they do not predict or grow nodes when there is insufficient data. Tree-based regression is known to be robust to monotonic behaviour of independent variables, so that the precise form in which these appear in the model is irrelevant. The standard linear model does not allow interactions between independent variables unless they are in multiplicative form. Tree-based models can detect interaction between parts of levels or parts of the numeric range of independent variables.

In growing a tree, the binary partitioning algorithm recursively splits the data in each node until either the node is homogeneous or the node contains too few observations. The minimum node deviance and the minimum number of observations in fitting a tree-based regression for small datasets varies according to the software and algorithms used. Robust method to determine the two mentioned criteria is still unknown. For huge dataset, it will result with inaccurate and brushy tree.

## 4.2 Construction of a Tree - Recursive Partitioning

In general, a regression tree model is fitted/trained on a training set using a binary recursive partitioning. In most cases the model follows Breiman et al. (1984) quite closely. A recursive partitioning package called ‘rpart’ was written for the software package R, a statistical programming environment. We utilise ‘rpart’ for our wool auction problem. This is an iterative process of splitting the data (the initial node) into binary partitions hence branching into two new nodes, and then splitting each new node further. Initially all of the records in the training set are together in the initial node. The algorithm then tries breaking up the data in the node, using every possible binary split on every input variable, splitting along coordinate axes of the these variables. From these splits, a single split is chosen to partition the data into two parts such that it minimises the sum of the squared deviations from the mean output in the separate parts.

$$D = \sum_{cases\, j} (y_j - \mu_j)^2 \qquad (4.2)$$

These two parts form the new nodes.

In the case of data of each wool auction sale lot there are six variables: DIAMETER, POBMID, SS, SL, VMB and YIELD. To illustrate the recursive partitioning, we consider the following example of wool dataset of size 10 (10 auction sale lots), whose output response CPRICE has a mean of 632.38:

Table 4.1: Example of Wool Auction Data of 10 Sale Lots

| | Input Variables (Wool Characteristics) | | | | | | Output Response |
|---|---|---|---|---|---|---|---|
| Sale Lot ID | DIAMETER | POBMID | SL | SS | VMB | YIELD | CPRICE |
| 1 | 20.1 | 64 | 73 | 29 | 1.2 | 59.8 | 602.01 |
| 2 | 20.2 | 53 | 65 | 37 | 2.0 | 52.0 | 576.92 |
| 3 | 19.0 | 57 | 72 | 26 | 4.3 | 54.6 | 961.54 |
| 4 | 20.3 | 59 | 77 | 26 | 0.8 | 70.5 | 567.38 |
| 5 | 21.8 | 60 | 76 | 26 | 1.2 | 55.8 | 498.21 |
| 6 | 23.1 | 38 | 84 | 22 | 1.9 | 54.0 | 461.11 |
| 7 | 20.4 | 37 | 70 | 32 | 3.7 | 47.3 | 505.29 |
| 8 | 19.7 | 49 | 65 | 34 | 2.2 | 56.4 | 721.63 |
| 9 | 23.1 | 50 | 87 | 36 | 3.0 | 55.4 | 476.53 |
| 10 | 18.9 | 61 | 64 | 33 | 2.4 | 59.8 | 953.18 |
| | | | | | | Mean: | 632.38 |

In this example we have 10 wool sale lots in the data, so the initial node is of size 10, with an average CPRICE of 632.38.

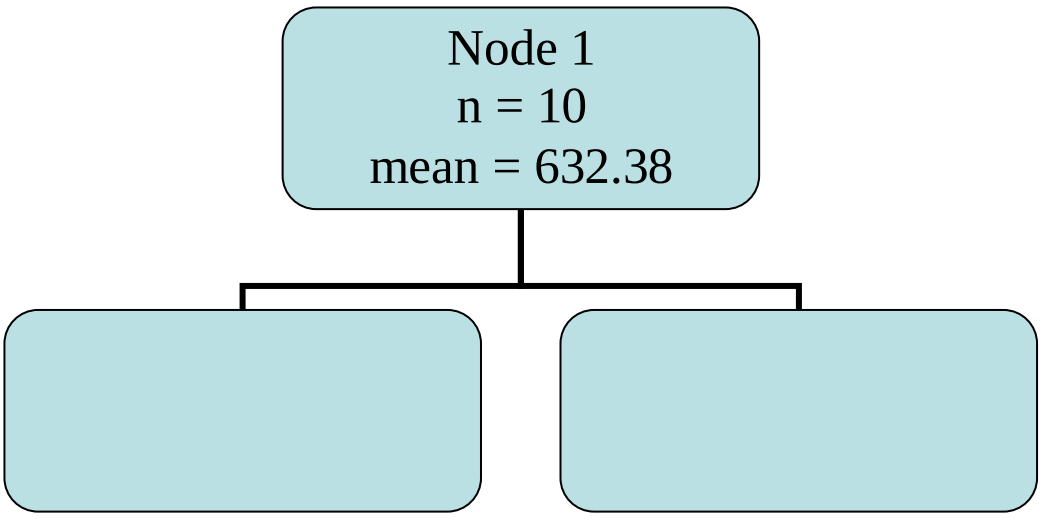


Figure 4.2: Initial Node

We then try breaking up the data in the node, using every possible binary split on every input variable, splitting along coordinate axes of the these variables. We consider the first variable DIAMETER. First we sort the data according to DIAMETER in an ascending order.

Table 4.2: Wool Auction Data Sorted according to DIAMETER

| | Input Variables (Wool Characteristics) | | | | | | Output Response |
|---|---|---|---|---|---|---|---|
| Sale Lot ID | **DIAMETER** | POBMID | SL | SS | VMB | YIELD | CPRICE |
| 10 | **18.9** | 61 | 64 | 33 | 2.4 | 59.8 | 953.18 |
| 3 | **19.0** | 57 | 72 | 26 | 4.3 | 54.6 | 961.54 |
| 8 | **19.7** | 49 | 65 | 34 | 2.2 | 56.4 | 721.63 |
| 1 | **20.1** | 64 | 73 | 29 | 1.2 | 59.8 | 602.01 |
| 2 | **20.2** | 53 | 65 | 37 | 2.0 | 52.0 | 576.92 |
| 4 | **20.3** | 59 | 77 | 26 | 0.8 | 70.5 | 567.38 |
| 7 | **20.4** | 37 | 70 | 32 | 3.7 | 47.3 | 505.29 |
| 5 | **21.8** | 60 | 76 | 26 | 1.2 | 55.8 | 498.21 |
| 6 | **23.1** | 38 | 84 | 22 | 1.9 | 54.0 | 461.11 |
| 9 | **23.1** | 50 | 87 | 36 | 3.0 | 55.4 | 476.53 |
| | | | | | | Mean: | 632.38 |

Here in this data of size 10 we have a total of 9 unique DIAMETERs. If we take the mid-point of DIAMETERs 18.9 and 19 as 18.95, and the mid-point of DIAMETERs 19 and 19.7 as 19.35 etc, there exists 8 possible mid-points (or splits) at DIAMETER: 18.95, 19.35, 19.9, 20.15, 20.25, 20.35, 21.1 and 22.45. If we take the first split at DIAMETER = 18.95 then the data would be divided into left and right as such:

Table 4.3: Splitting at DIAMETER = 18.95, part 1.

Left (DIAMETER < 18.95):

| | Input Variables (Wool Characteristics) | | | | | | Output Response |
|---|---|---|---|---|---|---|---|
| Sale Lot ID | DIAMETER | POBMID | SL | SS | VMB | YIELD | CPRICE |
| 10 | 18.9 | 61 | 64 | 33 | 2.4 | 59.8 | 953.18 |
| | | | | | | Mean: | 953.18 |

Right (DIAMETER > 18.95):

| | Input Variables (Wool Characteristics) | | | | | | Output Response |
|---|---|---|---|---|---|---|---|
| Sale Lot ID | DIAMETER | POBMID | SL | SS | VMB | YIELD | CPRICE |
| 3 | 19.0 | 57 | 72 | 26 | 4.3 | 54.6 | 961.54 |
| 8 | 19.7 | 49 | 65 | 34 | 2.2 | 56.4 | 721.63 |
| 1 | 20.1 | 64 | 73 | 29 | 1.2 | 59.8 | 602.01 |
| 2 | 20.2 | 53 | 65 | 37 | 2.0 | 52.0 | 576.92 |
| 4 | 20.3 | 59 | 77 | 26 | 0.8 | 70.5 | 567.38 |
| 7 | 20.4 | 37 | 70 | 32 | 3.7 | 47.3 | 505.29 |
| 5 | 21.8 | 60 | 76 | 26 | 1.2 | 55.8 | 498.21 |
| 6 | 23.1 | 38 | 84 | 22 | 1.9 | 54.0 | 461.11 |
| 9 | 23.1 | 50 | 87 | 36 | 3.0 | 55.4 | 476.53 |
| | | | | | | Mean: | 596.74 |

The mean outputs in the left and right partitions are noted: 953.18 and 596.74. Then the squared deviations of CPRICE from mean and their sum in each partition are calculated:

Table 4.4: Splitting at DIAMETER = 18.95, part 2.

Left (DIAMETER < 18.95):

| | Input Variables (Wool Characteristics) | | Output Response | Deviation from the left mean | Squared Deviation |
|---|---|---|---|---|---|
| Sale Lot ID | DIAMETER | … | CPRICE | 953.18 | |
| 10 | 18.9 | … | 953.18 | 0 | 0 |
| | | Mean: | 953.18 | Sum: | **0** |

Right (DIAMETER > 18.95):

| | Input Variables (Wool Characteristics) | | Output Response | Deviation from the right mean | Squared Deviation |
|---|---|---|---|---|---|
| Sale Lot ID | DIAMETER | … | CPRICE | 596.7356 | |
| 3 | 19.0 | | 961.54 | 364.8044 | 133082.2503 |
| 8 | 19.7 | | 721.63 | 124.8944 | 15598.61115 |
| 1 | 20.1 | | 602.01 | 5.2744 | 27.81929536 |
| 2 | 20.2 | | 576.92 | -19.8156 | 392.6580034 |
| 4 | 20.3 | … | 567.38 | -29.3556 | 861.7512514 |
| 7 | 20.4 | | 505.29 | -91.4456 | 8362.297759 |
| 5 | 21.8 | | 498.21 | -98.5256 | 9707.293855 |
| 6 | 23.1 | | 461.11 | -135.6256 | 18394.30338 |
| 9 | 23.1 | | 476.53 | -120.2056 | 14449.38627 |
| | | Mean: | 596.74 | Sum: | **200876.3712** |

The same procedure is then repeated for the next split at DIAMETER = 19.35 to get the left and right sums of squared deviations:

Table 4.5: Splitting at DIAMETER = 19.35

Left (DIAMETER < 19.35):

| | Input Variables (Wool Characteristics) | | Output Response | Deviation from the left mean | Squared Deviation |
|---|---|---|---|---|---|
| Sale Lot ID | DIAMETER | … | CPRICE | 957.36 | |
| 10 | 18.9 | … | 953.18 | -4.18 | 17.4724 |
| 3 | 19.0 | | 961.54 | 4.18 | 17.4724 |
| | | Mean: | 957.36 | Sum: | **34.9448** |

Right (DIAMETER > 19.35):

| | Input Variables (Wool Characteristics) | | Output Response | Deviation from the right mean | Squared Deviation |
|---|---|---|---|---|---|
| Sale Lot ID | DIAMETER | … | CPRICE | 551.135 | |
| 8 | 19.7 | | 721.63 | 170.495 | 29068.54503 |
| 1 | 20.1 | | 602.01 | 50.875 | 2588.265625 |
| 2 | 20.2 | | 576.92 | 25.785 | 664.866225 |
| 4 | 20.3 | … | 567.38 | 16.245 | 263.900025 |
| 7 | 20.4 | | 505.29 | -45.845 | 2101.764025 |
| 5 | 21.8 | | 498.21 | -52.925 | 2801.055625 |
| 6 | 23.1 | | 461.11 | -90.025 | 8104.500625 |
| 9 | 23.1 | | 476.53 | -74.605 | 5565.906025 |
| | | Mean: | 551.14 | Sum: | **51158.8032** |

And we repeat for all possible splits within the variable DIAMETER to get the following table:

Table 4.6: All possible splits with DIAMETER

| Split candidate | Mean CPRICE on left of split | Mean CPRICE on right of split | Sum of squared deviation of CPRICE from mean on left of split | Sum of squared deviation of CPRICE from mean on left of split | Sum of squared deviation on both sides | |
|---|---|---|---|---|---|---|
| 18.95 | 953.18 | 596.74 | 0 | 200876.4 | 200876.4 | |
| **19.35** | 957.36 | 551.14 | 34.9448 | 51158.8 | **51193.75** | **min** |
| 19.90 | 878.78 | 526.78 | 37080.7 | 17937.61 | 55018.31 | |
| 20.15 | 809.59 | 514.24 | 94533.31 | 11334.55 | 105867.9 | |
| 20.25 | 763.06 | 501.70 | 137841.6 | 6620.008 | 144461.6 | |
| 20.35 | 730.44 | 485.29 | 169749.2 | 1228.336 | 170977.5 | |
| 21.10 | 698.28 | 478.62 | 213201.2 | 694.7363 | 213895.9 | |
| 22.45 | 673.27 | 468.82 | 248225.2 | 118.8882 | 248344.1 | |

The split candidate 19.35 has the lowest corresponding sum of squared deviation at 51193.75, hence the best split for DIAMETER is 19.35.

After DIAMETER, we repeat the same procedure for POBMID (Table 4.7). This then leads us to Table 4.8.

Table 4.7: Wool Auction Data Sorted according to POBMID

| | Input Variables (Wool Characteristics) | | | | | | Output Response |
|---|---|---|---|---|---|---|---|
| Sale Lot ID | DIAMETER | **POBMID** | SL | SS | VMB | YIELD | CPRICE |
| 7 | 20.4 | **37** | 70 | 32 | 3.7 | 47.3 | 505.29 |
| 6 | 23.1 | **38** | 84 | 22 | 1.9 | 54.0 | 461.11 |
| 8 | 19.7 | **49** | 65 | 34 | 2.2 | 56.4 | 721.63 |
| 9 | 23.1 | **50** | 87 | 36 | 3.0 | 55.4 | 476.53 |
| 2 | 20.2 | **53** | 65 | 37 | 2.0 | 52.0 | 576.92 |
| 3 | 19.0 | **57** | 72 | 26 | 4.3 | 54.6 | 961.54 |
| 4 | 20.3 | **59** | 77 | 26 | 0.8 | 70.5 | 567.38 |
| 5 | 21.8 | **60** | 76 | 26 | 1.2 | 55.8 | 498.21 |
| 10 | 18.9 | **61** | 64 | 33 | 2.4 | 59.8 | 953.18 |
| 1 | 20.1 | **64** | 73 | 29 | 1.2 | 59.8 | 602.01 |
| | | | | | | Average: | 632.38 |

Table 4.8: All possible splits with POBMID

| Split candidate | Mean CPRICE on left of split | Mean CPRICE on right of split | Sum of squared deviation of CPRICE from mean on left of split | Sum of squared deviation of CPRICE from mean on left of split | Sum of squared deviation on both sides | |
|---|---|---|---|---|---|---|
| 37.5 | 505.29 | 646.50 | 0 | 297277.23 | 297277.23 | |
| 43.5 | 483.20 | 669.68 | 975.9362 | 258611.13 | 259587.07 | |
| 49.5 | 562.68 | 662.25 | 38875.179 | 255526.19 | 294401.37 | |
| 51.5 | 541.14 | 693.21 | 44441.116 | 215284.38 | 259725.5 | |
| **55.0** | 548.30 | 716.46 | 45465.282 | 199057.28 | **244522.56** | **min** |
| 58.0 | 617.17 | 655.20 | 187774.12 | 123979.47 | 311753.59 | |
| 59.5 | 610.06 | 684.47 | 189899.01 | 113697.5 | 303596.52 | |
| 60.5 | 596.08 | 777.60 | 200845.07 | 61660.184 | 262505.26 | |
| 62.5 | 635.75 | 602.01 | 314198.93 | 0 | 314198.93 | |

And we get the best split for POBMID which is at 55 with a sum of squared deviation at 244522.56. We repeat the procedure for all the remaining input variables (SL, SS, VMB and YIELD) and compare them:

Table 4.9: The best split with each input variable

| | Best Split | Sum of squared deviation | |
|---|---|---|---|
| DIAMETER | 19.35 | 51193.748 | optimal |
| POBMID | 55 | 244522.56 | |
| SL | 72.5 | 191275.61 | |
| SS | 24 | 282631.07 | |
| VMB | 4 | 194838.97 | |
| YIELD | 54.3 | 255610.13 | |

Splitting DIAMETER at 19.35 gives the minimum sum of squared deviation, hence it is the candidate that best partition the initial node so we choose it as our split for the initial node.

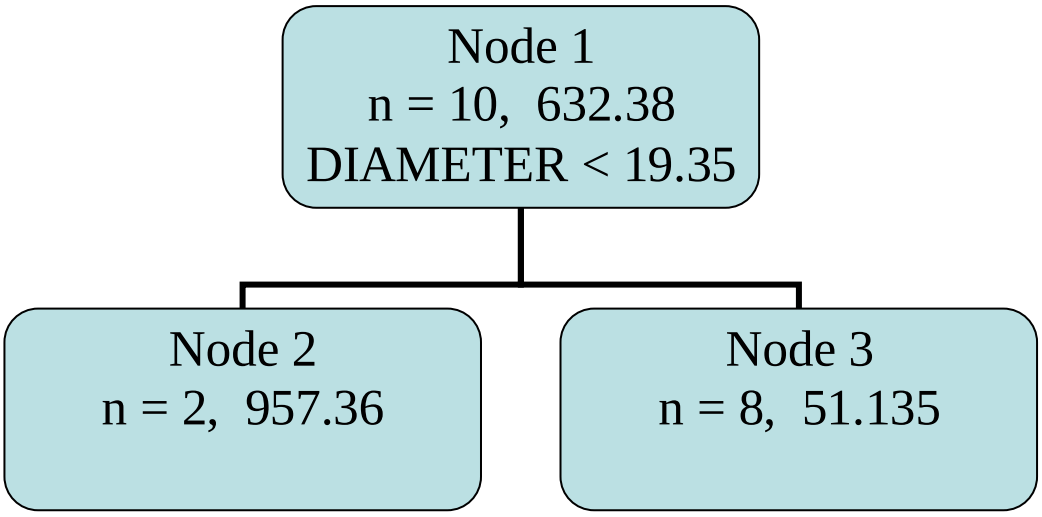


Figure 4.3: First split at DIAMETER = 19.35

Then, for each new node, the splitting procedure is performed exactly the same as for the initial node.

The process of splitting each node into new ones continues until nodes are pure (homogeneous), data are too sparse (too few observations), or each node reaches a user-specified minimum node size and becomes a terminal node. The minimum node deviance and the minimum number of observations in fitting a tree-based regression for small datasets varies according to the software and algorithm used. In the software package 'rpart', we can define the minimum

number of observations in a node before attempting a split. Robust methods to determine these criteria are still to be determined and are being investigated. The terminal nodes are called leaves, while the initial node is called the root. Since the response variable is numeric, the tree is called a regression tree. The model used for regression assumes that the numeric response variable has a normal (Gaussian) distribution.

### 4.3 Pruning of a Tree

With "noisy" data, it is quite possible to grow a tree which fits the training set well, but which has adapted too well to the features of this subset. Regression trees can be too elaborate and over-fit the training data. Say we have built a complete tree, possibly quite large and/or complex, and must now decide how much of that model to retain.

The established methodology is cost-complexity pruning, first introduced by Breiman et al. (1984). They considered rooted subtrees of a tree $T$ grown by the construction algorithm, that is the possible results of snipping off terminal subtrees on $T$. The pruning process chooses one of the rooted subtrees. Let $R_i$ be a measure evaluated at the leaves (terminal nodes), such as the deviance, and let $R$ be the value of the tree, the sum over the leaves of $R_i$. Let the size of the tree be the number of leaves i.e. define $|T|$ = number of terminal nodes.

Now let $\alpha$ be some number between 0 and $\infty$ which measures the "cost" of adding another variable to the model; $\alpha$ will be called a complexity parameter. Let $R(T_0)$ be the measure for the zero split tree. Define

$$R_\alpha(T) = R(T) + \alpha|T| \qquad (4.3)$$

to be the cost for the tree, and define $T_\alpha$ to be that subtree of the full model which has minimal cost. Obviously $T_0$ = the full model and $T_\infty$ = the model with no splits at all.

The following results are established by Breiman.

Result 1. If $T_1$ and $T_2$ are subtrees of $T$ with $R_\alpha(T_1) = R_\alpha(T_2)$, then either $T_1$ is a subtree of $T_2$ or $T_2$ is a subtree of $T_1$; hence either $|T_1| < |T_2|$ or $|T_2| < |T_1|$.

Result 2. If $\alpha > \beta$ then either $T_\alpha = T_\beta$ or $T_\alpha$ is a strict subtree of $T_\beta$.

Result 3. Given some set of numbers $\alpha_1, \alpha_2, \ldots, \alpha_m$; both $T_{\alpha_1}, T_{\alpha_2}, \ldots, T_{\alpha_m}$, and $R(T_{\alpha_1}), R(T_{\alpha_2}), \ldots, R(T_{\alpha_m})$ can be computed efficiently.

Using the first result, we can uniquely define $T_\alpha$ as the smallest tree $T$ for which $R_\alpha(T)$ is minimised.

Since any sequence of nested trees based on $T$ has at most $|T|$ members, Result 2 implies that all possible values of $\alpha$ can be grouped into $m$ intervals, $m \leq |T|$

$$\begin{aligned} I_1 &= [0, \alpha_1] \\ I_2 &= (\alpha_1, \alpha_2] \\ &\vdots \\ I_m &= (\alpha_{m-1}, \infty] \end{aligned} \tag{4.4}$$

where all $\alpha \in I_i$ share the same minimising subtree.

Breiman et al. (1984) showed that the set of rooted subtrees of $T$ which minimise the cost-complexity measure $R_\alpha(T)$ is itself nested. That is, as we increase $\alpha$ we can find the optimal trees by a sequence of snip operations on the current tree (just like pruning a real tree). This produces a sequence of trees from the size of $T$ down to just the root node, but it may prune more than one node at a time. The tree $T$ is not necessarily optimal for $\alpha = 0$.

We need a good way to choose the degree of pruning. If a separate validation set is available, we can predict on that set, and compute the deviance versus $\alpha$ for the pruned trees. This will often have a minimum, and we can choose the smallest tree whose deviance is close to the minimum.

If no validation set is available we can make one by splitting the training set. Suppose we split the training set into 10 (roughly) equally sized parts. We can then use 9 of these parts to grow the tree and use the remaining part to test this tree. This can be done in 10 ways, and we can average the results.

In actual practice, we may use instead the 1-SE rule. A plot of $\beta$ versus $R$ often has an initial sharp drop followed by a relatively flat plateau and then a slow rise. The choice of $\beta$ among those models on the plateau can be essentially random. To avoid this, both an estimate of $R$ and its standard error of the achieved minimum is marked as being equivalent to the minimum (i.e. considered to be part of the flat plateau). Then the simplest model, among all those "tied" on the plateau, is chosen.

In Monte-Carlo trials, this method of pruning has proven very reliable for screening out "pure noise" variables in the data set.

## 4.4 Additional Advantage of Regression Tree Over Neural Networks and Other Methods

In Section 1.3, we mentioned the existence of the additional wool specifications problem. Industrial sorting of wool during harvest, and at the start of processing, assembles wool in bins according to the required wool specifications. At present this assembly is done by constraining the range of all specifications in each bin, and having either a very large number of bins, or a large variance of characteristics within each bin. Neither neural networks nor older multiple linear regression on price could provide additional useful information that would streamline this process, nor did they assist in delineating the specifications of individual bins.

However, we have found a solution from the regression tree. Table 4.10 is derived directly from Figure 4.4 below, where the decisions (branching points in the regression tree) for each end point are extracted and condensed by removal of redundant decisions. The table shows results for a small but complete number of

bins, and for each bin, the ranges for fibre diameter (the most dominant characteristic of the raw wool) and supplementary characteristics where they are needed. The irregular occurrence of these supplementary terms, and the differences in the values at which they are used in the various bins demonstrate the complexity of the interactions that exist in the market.

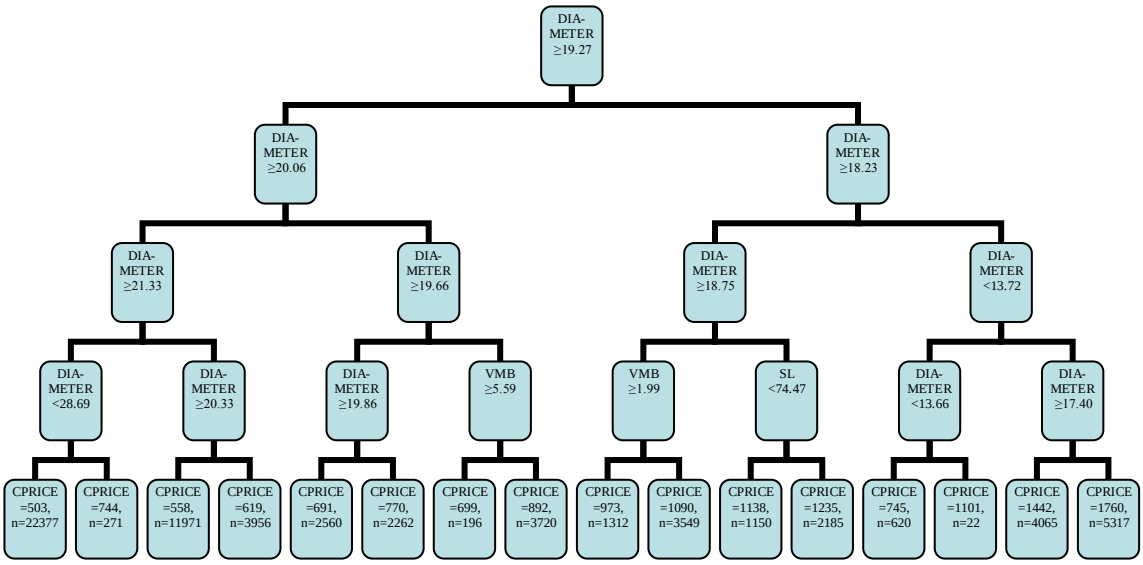


Figure 4.4: Another Example of a Regression Tree

Table 4.10: Fitting for Period A with Regression Tree

| DIAMETER min | DIAMETER Max | POBMID Min | POBMID max | SL min | SL max | SS min | SS max | VMB min | VMB max | YIELD min | YIELD max | PRICE Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 13.72 | 17.40 | | | | | | | | | | | 1760 |
| 17.40 | 18.23 | | | | | | | | | | | 1442 |
| 18.23 | 18.75 | | | 74.47 | | | | | | | | 1235 |
| 18.23 | 18.75 | | | | 74.47 | | | | | | | 1138 |
| 13.66 | 13.72 | | | | | | | | | | | 1101 |
| 18.75 | 19.27 | | | | | | | | 1.99 | | | 1090 |
| 18.75 | 19.27 | | | | | | | 1.99 | | | | 973 |
| 19.27 | 19.66 | | | | | | | | 5.59 | | | 892 |
| 19.66 | 19.86 | | | | | | | | | | | 770 |
| | 13.66 | | | | | | | | | | | 745 |
| 28.69 | | | | | | | | | | | | 744 |
| 19.27 | 19.66 | | | | | | | 5.59 | | | | 699 |
| 19.86 | 20.06 | | | | | | | | | | | 691 |
| 20.06 | 20.33 | | | | | | | | | | | 619 |
| 20.33 | 21.33 | | | | | | | | | | | 558 |
| 21.33 | 28.69 | | | | | | | | | | | 503 |

For example, from Table 4.10 we can conclude that during the period of this table, wool with DIAMETER between 18.23 and 18.75 and SL greater than 74.47 would give a price of about 1235, or vice versa. On the other hand, wool with DIAMETER between 19.27 and 19.66 and VMB less than 5.59 would give

a price of about 892. Thus the wool growers can use this as a guideline when assembling their wool into bins.

The algorithm to achieve this is quite simple and we illustrate this with Figure 4.4. The tree diagram has 16 end nodes, thus representing the average prices of 16 different groups of wool with similar wool specifications in each group (i.e. 16 different price levels). Say we accept that the 16 average prices represent a good variation in price levels and hence we would like to have the same partition in our table representation. Then we can backtrack from each end node. For example, say we wish to find the wool specifications for the end node with CPRICE = 1235. Backtracking from this end node to the top of the tree (Figure 4.4b) and we obtain the following information: SL>74.47, DIAMETER<18.75, DIAMETER>18.23 and DIAMETER<19.27. Combining the information gives us the summary that this particular group of wool has a DIAMETER between 18.23 and 18.75, and SL greater than 74.47. Thus we can then repeat the backtracking for each end node, and come up with the table shown in Table 4.10.

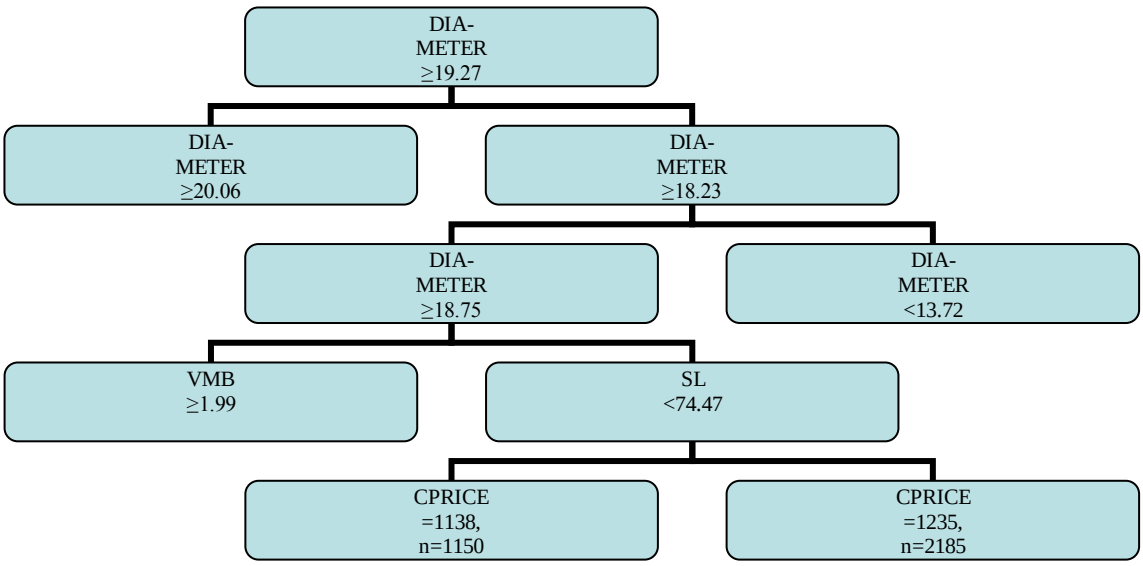


Figure 4.4b: Back Tracking from end node with CPRICE=1235

The number of CPRICE levels in the table (or in other words, how finely we wish to partition the CPRICE into different levels) depends on the number of end nodes in the tree diagram. Thus, if we wish to have a finer partitioning with many CPRICE levels, we can continue the recursive partitioning procedure on

each end node as detailed in Section 4.2, and grow a larger tree with many more end nodes. One the other hand, if we wish to have a smaller table with less CPRICE levels, then we can do the opposite by pruning the tree to make it smaller with a smaller number of end nodes.

So we can now come up with tabular representations of tree diagrams, which are useful in streamlining the process of assembling wool into bins and assist in delineating the specifications of individual bins.

## 4.5 Applying Regression Tree to the Wool Auction Data

In this section we assess regression tree's ability to model auction data and predict wool prices. All trees were generated using the 'rpart' package developed for the R statistical computation environment. The package was originally written for the S-PLUS statistical package by Terry M Therneau and Beth Atkinson. It was later ported to R by Brian Ripley. 'rpart' follows the algorithm of Breiman et al. (1984) quite closely. We use 'rpart' to generate a tree model and we store this model in memory. We assume 'CPRICE' to be some unknown function of 'DIAMETER', 'POBMID', 'SL', 'SS', 'VMB' and 'YIELD', as described in Section 2.2. We can then use the tree model stored in memory to predict the outcome for another dataset. As mentioned in Section 4.2, In 'rpart', we can define the minimum number of observations in a node before attempting a split. But since we can always prune back a large tree using cost-complexity pruning, we can always allow a tree to grow to its furthest reach then prune it afterwards. So we simply set 2 as the minimum number of observations in a node before attempting a split.

Again, we use the three periods introduced in Section 2.3 to assess regression tree's ability to model (or fit) the data, as well as the accuracy in prediction. We model the last week worth of auction data from each month, then use the model to predict the price outcomes in the first week of the following month. Our results follow.

On the following pages, Tables 4.11, 4.12 and 4.13 show the results from modelling (fitting) the last week of each month in Periods A, B and C with regression tree, while Tables 4.14, 4.15 and 4.16 show the results from predicting the last week of each month with those fitted models. We find that regression tree, with its advantage over neural networks (Section 3.4), unfortunately offers much poorer fitting and prediction accuracies than neural networks. This is a rather large problem that needs to be addressed. We also find that conventional pruning method offers no advantage in improving predictions, in the case of our data. We will consider some methods that can improve the accuracies in the next section.

Table 4.11: Fitting for Period A with Regression Tree

| | | GRNN | Regression Tree (without pruning) | Regression Tree (with pruning) |
|---|---|---|---|---|
| Fitting last week of Aug 2000 | Root Mean Square Error | 28.29091354 | 44.50163 | 48.94125 |
| | Mean Absolute Error | 15.26685036 | 23.3963 | 29.59593 |
| | Std. Deviation of Abs. Error | 23.82596829 | 37.86769 | 38.99152 |
| Fitting last week of Sep 2000 | Root Mean Square Error | 33.47677404 | 43.89331 | 50.47487 |
| | Mean Absolute Error | 16.33235456 | 23.13326 | 31.38879 |
| | Std. Deviation of Abs. Error | 29.23013271 | 37.31235 | 39.53838 |
| Fitting last week of Oct 2000 | Root Mean Square Error | 35.59787003 | 60.89278 | 75.55599 |
| | Mean Absolute Error | 20.77791974 | 29.00816 | 46.43321 |
| | Std. Deviation of Abs. Error | 28.9127874 | 53.55414 | 59.62074 |
| Fitting last week of Nov 2000 | Root Mean Square Error | 53.14404304 | 62.8968 | 90.33876 |
| | Mean Absolute Error | 17.53775253 | 30.5585 | 54.50748 |
| | Std. Deviation of Abs. Error | 50.1885246 | 54.99812 | 72.07291 |

Table 4.12: Fitting for Period B with Regression Tree

| | | GRNN | Regression Tree (without pruning) | Regression Tree (with pruning) |
|---|---|---|---|---|
| Fitting last week of Aug 2001 | Root Mean Square Error | 35.52785964 | 50.75957 | 58.57987 |
| | Mean Absolute Error | 19.05397627 | 27.20509 | 34.81702 |
| | Std. Deviation of Abs. Error | 29.99838122 | 42.87078 | 47.12932 |
| Fitting last week of Sep 2001 | Root Mean Square Error | 43.38224682 | 49.15919 | 61.59324 |
| | Mean Absolute Error | 18.91764899 | 21.53069 | 34.98672 |
| | Std. Deviation of Abs. Error | 39.04907165 | 44.20336 | 50.70322 |
| Fitting last week of Oct 2001 | Root Mean Square Error | 29.53830859 | 35.49992 | 46.65728 |
| | Mean Absolute Error | 16.2291678 | 19.29468 | 28.82508 |
| | Std. Deviation of Abs. Error | 24.6922348 | 29.81285 | 36.70558 |
| Fitting last week of Nov 2001 | Root Mean Square Error | 21.18334939 | 35.87093 | 44.41908 |
| | Mean Absolute Error | 11.16225224 | 19.22255 | 27.81427 |
| | Std. Deviation of Abs. Error | 18.01361591 | 30.30203 | 34.65145 |

Table 4.13: Fitting for Period C with Regression Tree

| | | GRNN | Regression Tree (without pruning) | Regression Tree (with pruning) |
|---|---|---|---|---|
| Fitting last week of Aug 2002 | Root Mean Square Error | 27.4876211 | 44.38823 | 53.55512 |
| | Mean Absolute Error | 15.47762742 | 23.93032 | 31.71296 |
| | Std. Deviation of Abs. Error | 22.72436108 | 37.39913 | 43.17204 |
| Fitting last week of Sep 2002 | Root Mean Square Error | 39.05447769 | 58.038 | 72.75251 |
| | Mean Absolute Error | 25.27669893 | 33.16222 | 46.31056 |
| | Std. Deviation of Abs. Error | 29.78166889 | 47.64693 | 56.12857 |
| Fitting last week of Oct 2002 | Root Mean Square Error | 33.21652489 | 39.58191 | 47.87815 |
| | Mean Absolute Error | 19.23416188 | 24.98678 | 31.93046 |
| | Std. Deviation of Abs. Error | 27.09155177 | 30.71022 | 35.68961 |
| Fitting last week of Nov 2002 | Root Mean Square Error | 33.74128347 | 38.34854 | 42.66919 |
| | Mean Absolute Error | 21.64586617 | 23.98833 | 29.06492 |
| | Std. Deviation of Abs. Error | 25.89736289 | 29.93598 | 31.25656 |

Table 4.14: Predictions for Period A with Regression Tree

| | | GRNN | Regression Tree (without pruning) | Regression Tree (with pruning) |
|---|---|---|---|---|
| Using last week of Aug 2000 to predict 1st wk of Sep 2000 | Root Mean Square Error | 47.88411498 | 65.63687 | 68.37046 |
| | Mean Absolute Error | 27.49475198 | 34.04639 | 38.86008 |
| | Std. Deviation of Abs. Error | 39.21427868 | 56.13152 | 56.26836 |
| Using last week of Sep 2000 to predict 1st wk of Oct 2000 | Root Mean Square Error | 135.7182532 | 136.596 | 137.7818 |
| | Mean Absolute Error | 29.63332056 | 36.90987 | 41.17625 |
| | Std. Deviation of Abs. Error | 132.4945037 | 131.5653 | 131.5357 |
| Using last week of Oct 2000 to predict 1st wk of Nov 2000 | Root Mean Square Error | 69.12056444 | 96.05168 | 102.1285 |
| | Mean Absolute Error | 35.6166384 | 48.66818 | 58.60488 |
| | Std. Deviation of Abs. Error | 59.2514565 | 82.82823 | 83.65973 |
| Using last week of Nov 2000 to predict 1st wk of Dec 2000 | Root Mean Square Error | 64.7849604 | 83.60047 | 104.3844 |
| | Mean Absolute Error | 34.43145383 | 45.3864 | 63.93187 |
| | Std. Deviation of Abs. Error | 54.90408942 | 70.24136 | 82.5552 |

Table 4.15: Predictions for Period B with Regression Tree

| | | GRNN | Regression Tree (without pruning) | Regression Tree (with pruning) |
|---|---|---|---|---|
| Using last week of Aug 2001 to predict 1st wk of Sep 2001 | Root Mean Square Error | 58.66584341 | 69.7189 | 72.83096 |
| | Mean Absolute Error | 32.43893849 | 36.92937 | 40.31417 |
| | Std. Deviation of Abs. Error | 48.8988455 | 59.15603 | 60.6773 |
| Using last week of Sep 2001 to predict 1st wk of Oct 2001 | Root Mean Square Error | 89.2013982 | 93.72725 | 101.1088 |
| | Mean Absolute Error | 74.55923056 | 73.82104 | 79.49163 |
| | Std. Deviation of Abs. Error | 48.97674904 | 57.76261 | 62.49446 |
| Using last week of Oct 2001 to predict 1st wk of Nov 2001 | Root Mean Square Error | 38.35475621 | 44.31331 | 48.19521 |
| | Mean Absolute Error | 27.41881763 | 30.44673 | 31.49233 |
| | Std. Deviation of Abs. Error | 26.82817574 | 32.20748 | 36.49457 |
| Using last week of Nov 2001 to predict 1st wk of Dec 2001 | Root Mean Square Error | 66.14964581 | 71.31696 | 76.90712 |
| | Mean Absolute Error | 54.72642547 | 57.22881 | 60.07254 |
| | Std. Deviation of Abs. Error | 37.18058425 | 42.5802 | 48.04862 |

Table 4.16: Predictions for Period C with Regression Tree

| | | GRNN | Regression Tree (without pruning) | Regression Tree (with pruning) |
|---|---|---|---|---|
| Using last week of Aug 2002 to predict 1st wk of Sep 2002 | Root Mean Square Error | 38.60712508 | 49.9932 | 53.07456 |
| | Mean Absolute Error | 27.09162225 | 32.91185 | 35.77446 |
| | Std. Deviation of Abs. Error | 27.51594552 | 37.64576 | 39.22068 |
| Using last week of Sep 2002 to predict 1st wk of Oct 2002 | Root Mean Square Error | 67.49464254 | 77.1083 | 83.63511 |
| | Mean Absolute Error | 42.78633291 | 49.20164 | 55.48768 |
| | Std. Deviation of Abs. Error | 52.20917584 | 59.38102 | 62.58836 |
| Using last week of Oct 2002 to predict 1st wk of Nov 2002 | Root Mean Square Error | 53.65067084 | 60.417 | 62.70145 |
| | Mean Absolute Error | 38.96359524 | 43.58813 | 45.70082 |
| | Std. Deviation of Abs. Error | 36.89281363 | 41.84948 | 42.94245 |
| Using last week of Nov 2002 to predict 1st wk of Dec 2002 | Root Mean Square Error | 41.68142591 | 46.99935 | 48.9866 |
| | Mean Absolute Error | 28.36965042 | 32.1573 | 34.00197 |
| | Std. Deviation of Abs. Error | 30.54913432 | 34.28974 | 35.27815 |

## 4.6 Ensemble Methods

To improve both the fitting and the prediction results from regression trees, some authors (Breiman 1996, 1998, 1999, 2000, 2001) have observed that combining a multiple set of predictors, all constructed using the same data, can lead to dramatic decreases in test error. One of the most promising methods is bagging (as in bootstrap aggregating) (Breiman 1996). In its application, we take multiple bootstrap samples from the learning set of data (sampling with replacement), then grow a tree from each bootstrap sample. The predictions from these trees can then be averaged to give us an overall prediction, which is a much improved prediction than a single ordinary regression tree. In essence, the bootstrap procedure introduced randomness which reduced variance in the data, so the model built from it can be more accurate.

The study of bagging then led to random forests (Ho 1995, Breiman 2001) which introduced even more randomness in the process. When generating a tree from each bootstrap sample, instead of choosing an optimal split from all variables at each node (Section 4.2), only a random selection of the variables is considered. In essence, this refinement improves on bagging by "de-correlating" the trees. Other variants of these methods also exist and they are all based on the same model averaging approach (Barutcuoglu and Alpaydin 2003; Breiman 1996, 1998, 1999; Frank and Pfahringer 2006; Freund and Schapire 1995; Friedman et al 2000; Friedman 2001). Figure 4.5 on the following page shows a summary of the ensemble methods in general.

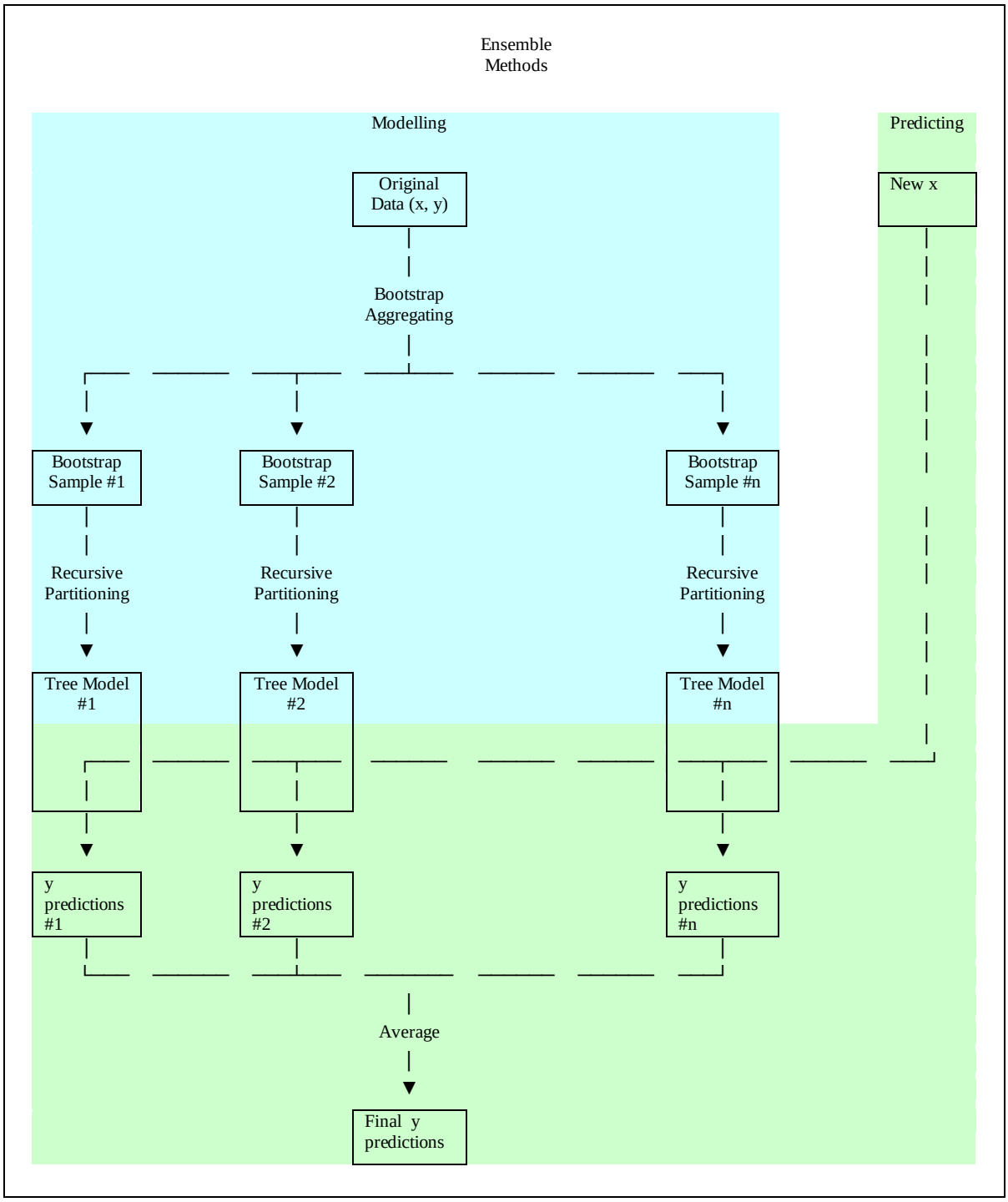


Figure 4.5: Diagrammatic Summary of Ensemble Methods

In the tables and plots that follow, we present the alternate results from bagging and random forests, and compare them with the earlier results from both GRNN and regression tree. Bagging was performed using the 'bagging' package and the

random forests were generated using the 'randomForest' package. Both software packages, just like the earlier 'rpart' package, are also packages developed for the R statistical computation environment. The number of trees used in each procedure of bagging was 25 (default setting), while the number of trees used in each random forest was 100 by default.

We observe that both bagging and random forests are in general, relatively poor in fitting the data, even poorer than simple regression tree. This is possibly due to the regression tree over-fitting the data and showing better numbers. However, bagging really shines through when we look at the predictions. While the results from fitting are poor, the results from predictions are on par with those from GRNN in Periods A, B and C. Random forests, on the other hand, give poorer results than bagging in general. For our wool auction data, we find random forests to be not as computational economical as bagging, considering that 25 trees were used in bagging in each case but 100 trees were used in random forests.

Another popular ensemble method in the literature: "boosting" (Friedman et al 2000) would also be a logical candidate for considerations after bagging and random forests. However, the boosting packages from R appear to perform poorly with the wool data that is available to us. It would be computational expensive for boosting to achieve similar results to bagging and random forests. The number of trees required is impractical when compared to other methods, thus we decided to omit boosting in our comparisons.

Tables 4.17, 4.18 and 4.19 show the results in fitting using Ensemble Methods:

Table 4.17: Fitting for Period A with Ensemble Methods

| | | GRNN | Regression Tree (without pruning) | Bagging | Random Forest |
|---|---|---|---|---|---|
| Fitting last week of Aug 2000 | Root Mean Square Error | 28.29091354 | 44.50163 | 50.39427 | 52.97568 |
| | Mean Absolute Error | 15.26685036 | 23.3963 | 25.70137 | 28.17591 |
| | Std. Deviation of Abs. Error | 23.82596829 | 37.86769 | 43.36212 | 44.87628 |
| Fitting last week of Sep 2000 | Root Mean Square Error | 33.47677404 | 43.89331 | 50.58865 | 59.96171 |
| | Mean Absolute Error | 16.33235456 | 23.13326 | 26.0236 | 31.0208 |
| | Std. Deviation of Abs. Error | 29.23013271 | 37.31235 | 43.39332 | 51.32748 |
| Fitting last week of Oct 2000 | Root Mean Square Error | 35.59787003 | 60.89278 | 64.65287 | 76.3059 |
| | Mean Absolute Error | 20.77791974 | 29.00816 | 32.8266 | 37.83955 |
| | Std. Deviation of Abs. Error | 28.9127874 | 53.55414 | 55.71469 | 66.28115 |
| Fitting last week of Nov 2000 | Root Mean Square Error | 53.14404304 | 62.8968 | 82.35355 | 82.01819 |
| | Mean Absolute Error | 17.53775253 | 30.5585 | 32.94694 | 40.16307 |
| | Std. Deviation of Abs. Error | 50.1885246 | 54.99812 | 75.50842 | 71.54246 |

Table 4.18: Fitting for Period B with Ensemble Methods

| | | GRNN | Regression Tree (without pruning) | Bagging | Random Forest |
|---|---|---|---|---|---|
| Fitting last week of Aug 2001 | Root Mean Square Error | 35.52785964 | 50.75957 | 58.01292 | 59.49253 |
| | Mean Absolute Error | 19.05397627 | 27.20509 | 31.84648 | 33.15729 |
| | Std. Deviation of Abs. Error | 29.99838122 | 42.87078 | 48.50984 | 49.4159 |
| Fitting last week of Sep 2001 | Root Mean Square Error | 43.38224682 | 49.15919 | 58.10737 | 57.9892 |
| | Mean Absolute Error | 18.91764899 | 21.53069 | 27.37857 | 27.32785 |
| | Std. Deviation of Abs. Error | 39.04907165 | 44.20336 | 51.26467 | 51.15776 |
| Fitting last week of Oct 2001 | Root Mean Square Error | 29.53830859 | 35.49992 | 44.13764 | 41.61895 |
| | Mean Absolute Error | 16.2291678 | 19.29468 | 24.34452 | 23.20486 |
| | Std. Deviation of Abs. Error | 24.6922348 | 29.81285 | 36.83433 | 34.56602 |
| Fitting last week of Nov 2001 | Root Mean Square Error | 21.18334939 | 35.87093 | 42.29404 | 44.65671 |
| | Mean Absolute Error | 11.16225224 | 19.22255 | 22.56911 | 22.9732 |
| | Std. Deviation of Abs. Error | 18.01361591 | 30.30203 | 35.78841 | 38.31509 |

Table 4.19: Fitting for Period C with Ensemble Methods

| | | GRNN | Regression Tree (without pruning) | Bagging | Random Forest |
|---|---|---|---|---|---|
| Fitting last week of Aug 2002 | Root Mean Square Error | 27.4876211 | 44.38823 | 54.82166 | 47.84726 |
| | Mean Absolute Error | 15.47762742 | 23.93032 | 28.40097 | 26.17418 |
| | Std. Deviation of Abs. Error | 22.72436108 | 37.39913 | 46.90882 | 40.06829 |
| Fitting last week of Sep 2002 | Root Mean Square Error | 39.05447769 | 58.038 | 67.73973 | 65.07719 |
| | Mean Absolute Error | 25.27669893 | 33.16222 | 40.58841 | 39.37019 |
| | Std. Deviation of Abs. Error | 29.78166889 | 47.64693 | 54.25188 | 51.835 |
| Fitting last week of Oct 2002 | Root Mean Square Error | 33.21652489 | 39.58191 | 48.19551 | 47.27497 |
| | Mean Absolute Error | 19.23416188 | 24.98678 | 30.47481 | 29.50627 |
| | Std. Deviation of Abs. Error | 27.09155177 | 30.71022 | 37.352 | 36.95077 |
| Fitting last week of Nov 2002 | Root Mean Square Error | 33.74128347 | 38.34854 | 46.43959 | 44.55623 |
| | Mean Absolute Error | 21.64586617 | 23.98833 | 28.34224 | 27.48117 |
| | Std. Deviation of Abs. Error | 25.89736289 | 29.93598 | 36.80834 | 35.0914 |

Tables 4.20, 4.21 and 4.22, and Figures 4.6 to 4.29 show the results in predicting using Ensemble Methods:

Table 4.20: Predictions for Period A with Ensemble Methods

| | | GRNN | Regression Tree (without pruning) | Bagging | Random Forest |
|---|---|---|---|---|---|
| Using last week of Aug 2000 to predict 1st wk of Sep 2000 | Root Mean Square Error | 47.88411498 | 65.63687 | 50.9104 | 60.12618 |
| | Mean Absolute Error | 27.49475198 | 34.04639 | 26.91575 | 30.97925 |
| | Std. Deviation of Abs. Error | 39.21427868 | 56.13152 | 43.22526 | 51.54495 |
| Using last week of Sep 2000 to predict 1st wk of Oct 2000 | Root Mean Square Error | 135.7182532 | 136.596 | 132.3349 | 148.8892 |
| | Mean Absolute Error | 29.63332056 | 36.90987 | 29.8902 | 36.99417 |
| | Std. Deviation of Abs. Error | 132.4945037 | 131.5653 | 128.9646 | 144.2755 |
| Using last week of Oct 2000 to predict 1st wk of Nov 2000 | Root Mean Square Error | 69.12056444 | 96.05168 | 74.725 | 84.0915 |
| | Mean Absolute Error | 35.6166384 | 48.66818 | 37.72461 | 45.26432 |
| | Std. Deviation of Abs. Error | 59.2514565 | 82.82823 | 64.51828 | 70.88619 |
| Using last week of Nov 2000 to predict 1st wk of Dec 2000 | Root Mean Square Error | 64.7849604 | 83.60047 | 69.56963 | 94.32711 |
| | Mean Absolute Error | 34.43145383 | 45.3864 | 37.52524 | 44.5789 |
| | Std. Deviation of Abs. Error | 54.90408942 | 70.24136 | 58.60961 | 83.16829 |

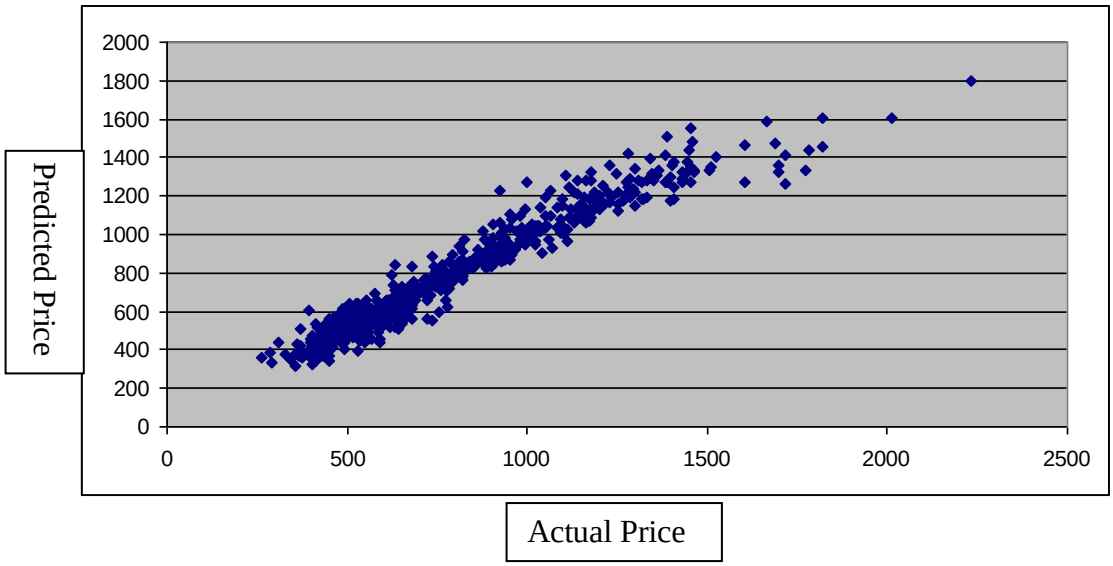


Figure 4.6: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of September 2000 with Bagging

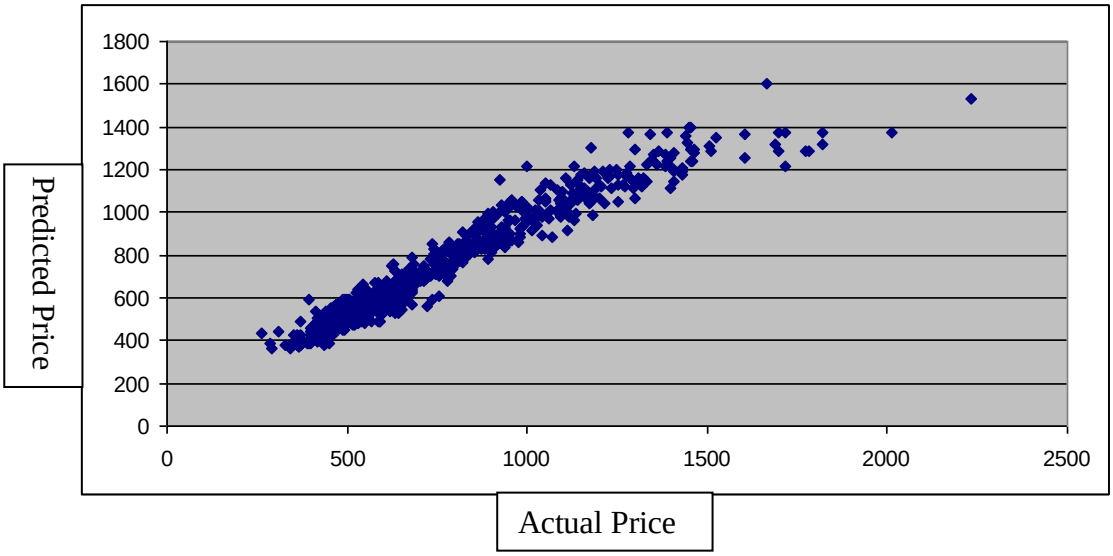


Figure 4.7: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of September 2000 with Random Forest

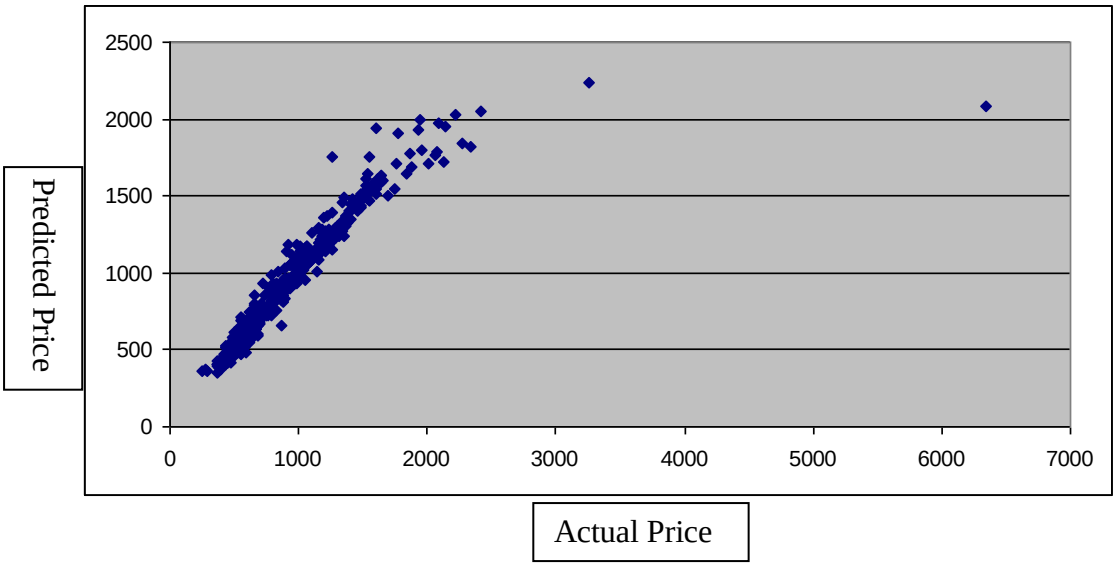


Figure 4.8: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of October 2000 with Bagging

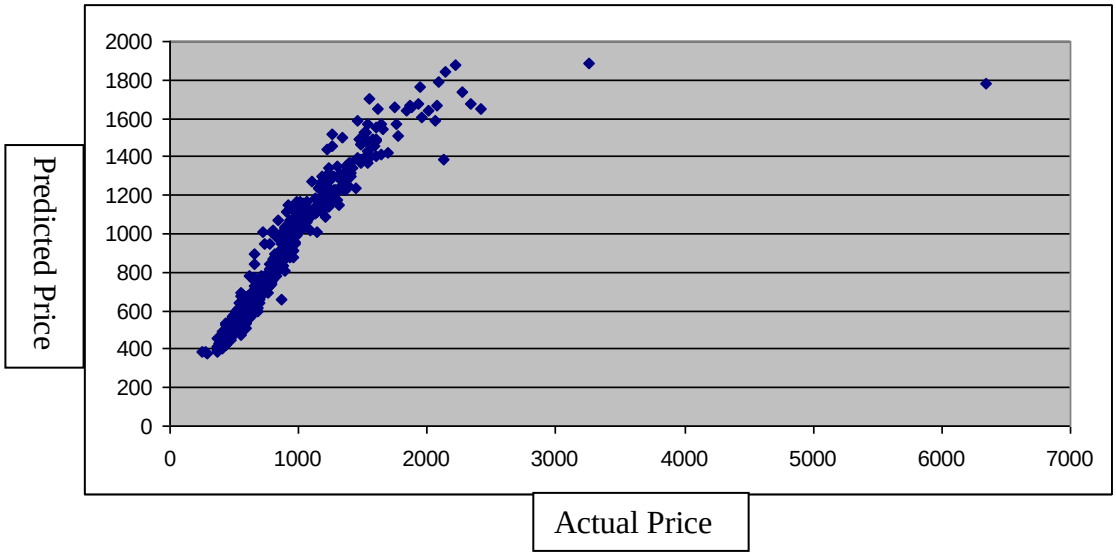


Figure 4.9: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of October 2000 with Random Forest

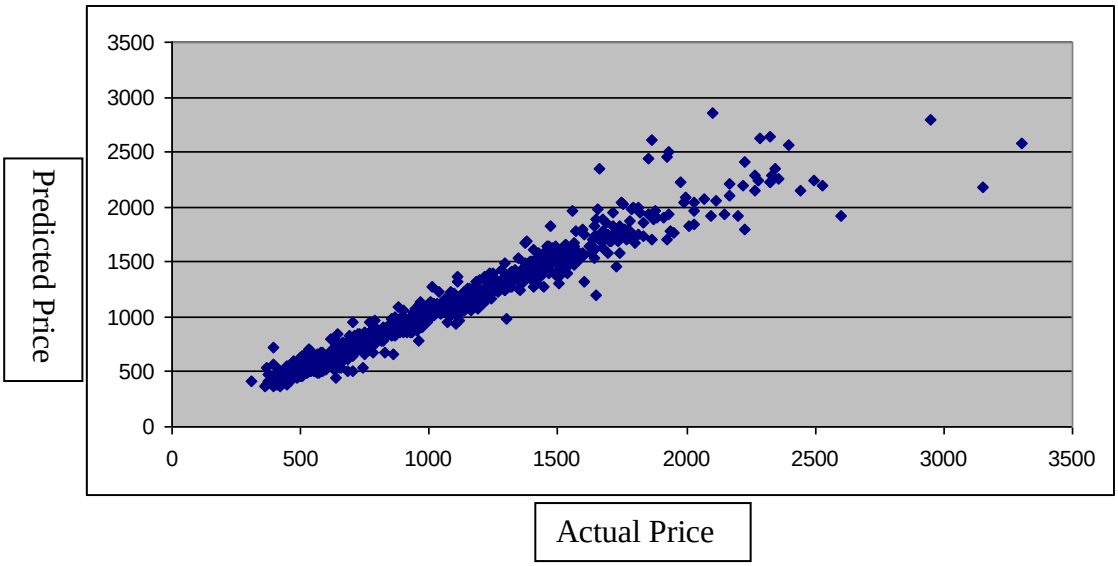


Figure 4.10: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of November 2000 with Bagging

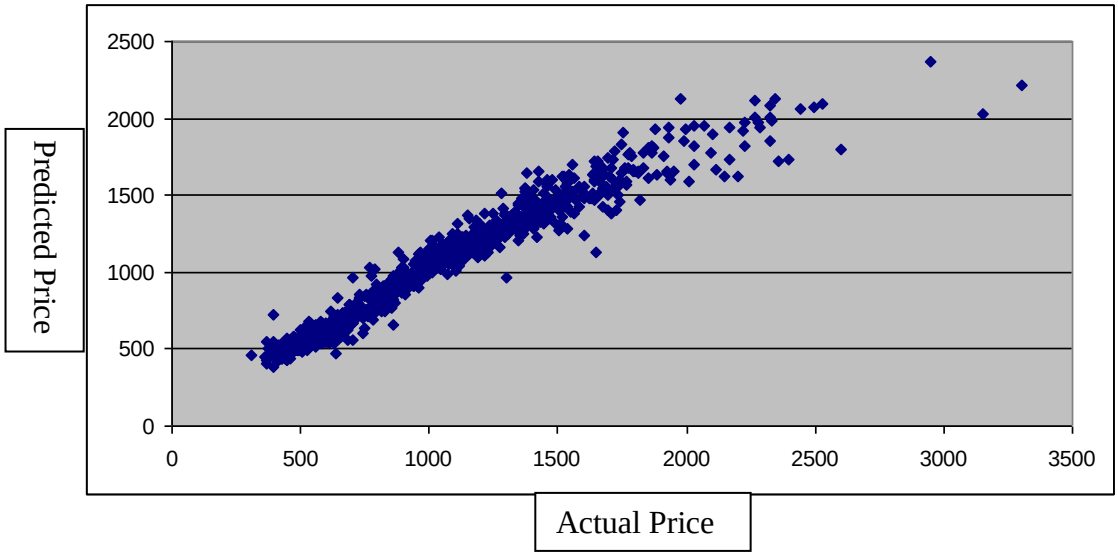


Figure 4.11: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of November 2000 with Random Forest

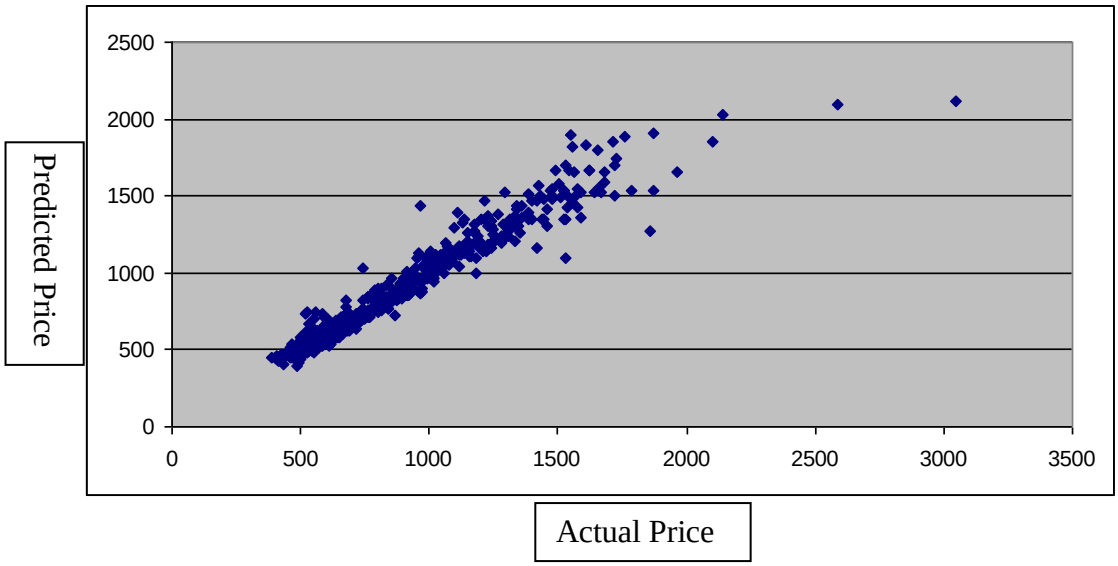


Figure 4.12: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of December 2000 with Bagging

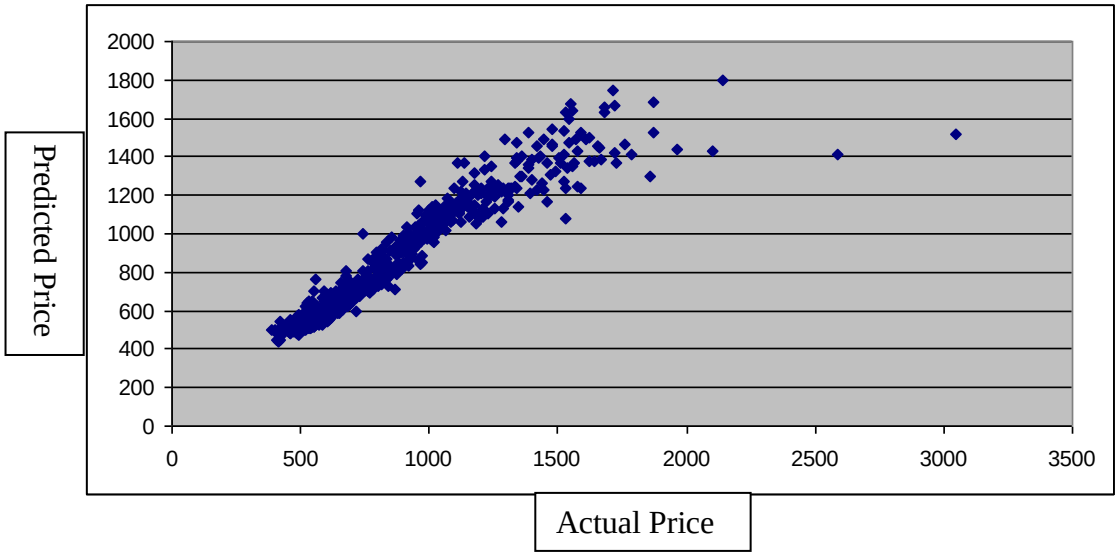


Figure 4.13: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of December 2000 with Random Forest

Table 4.21: Predictions for Period B with Ensemble Methods

| | | GRNN | Regression Tree (without pruning) | Bagging | Random Forest |
|---|---|---|---|---|---|
| Using last week of Aug 2001 to predict 1st wk of Sep 2001 | Root Mean Square Error | 58.66584341 | 69.7189 | 58.85238 | 63.08733 |
| | Mean Absolute Error | 32.43893849 | 36.92937 | 30.61869 | 32.26526 |
| | Std. Deviation of Abs. Error | 48.8988455 | 59.15603 | 50.27818 | 54.23151 |
| Using last week of Sep 2001 to predict 1st wk of Oct 2001 | Root Mean Square Error | 89.2013982 | 93.72725 | 88.41392 | 85.0301 |
| | Mean Absolute Error | 74.55923056 | 73.82104 | 72.65839 | 72.55067 |
| | Std. Deviation of Abs. Error | 48.97674904 | 57.76261 | 50.38595 | 44.35385 |
| Using last week of Oct 2001 to predict 1st wk of Nov 2001 | Root Mean Square Error | 38.35475621 | 44.31331 | 39.72906 | 38.51702 |
| | Mean Absolute Error | 27.41881763 | 30.44673 | 26.51407 | 26.55256 |
| | Std. Deviation of Abs. Error | 26.82817574 | 32.20748 | 29.59656 | 27.91083 |
| Using last week of Nov 2001 to predict 1st wk of Dec 2001 | Root Mean Square Error | 66.14964581 | 71.31696 | 66.97028 | 67.96612 |
| | Mean Absolute Error | 54.72642547 | 57.22881 | 55.07718 | 52.90601 |
| | Std. Deviation of Abs. Error | 37.18058425 | 42.5802 | 38.12095 | 42.69028 |

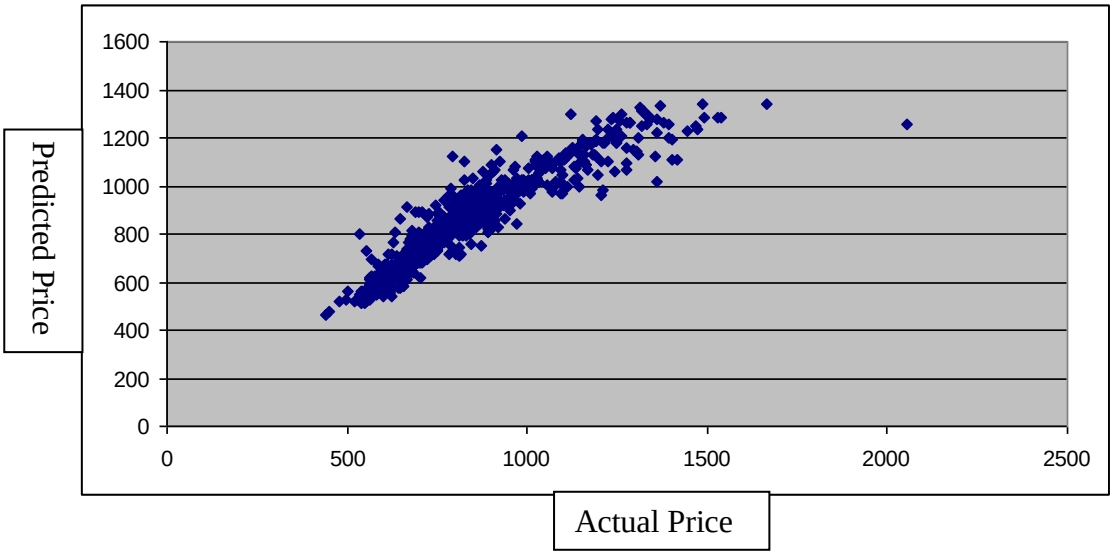


Figure 4.14: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of September 2001 with Bagging

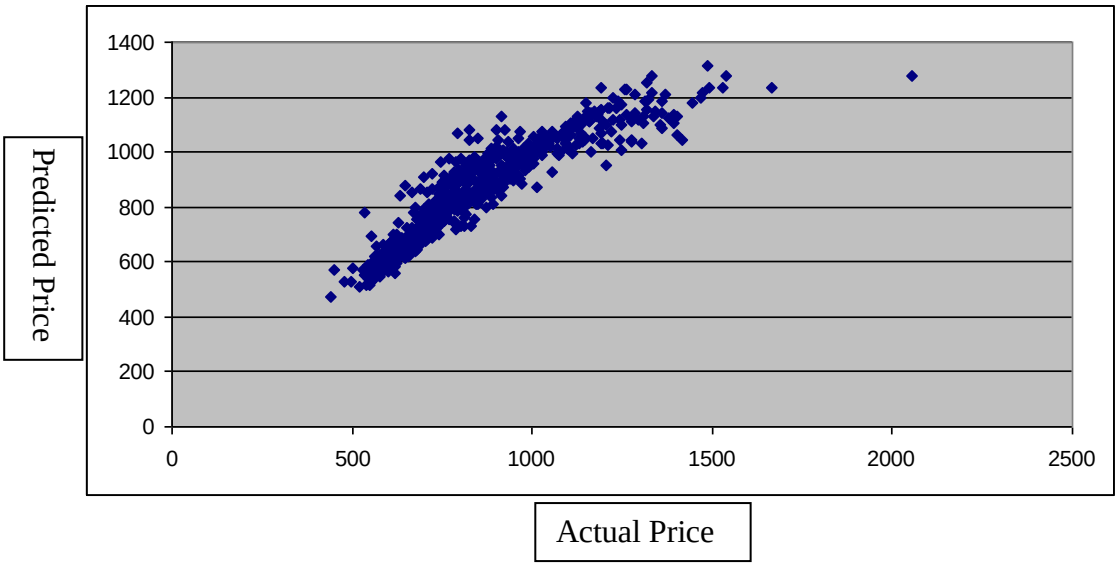


Figure 4.15: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of September 2001 with Random Forest

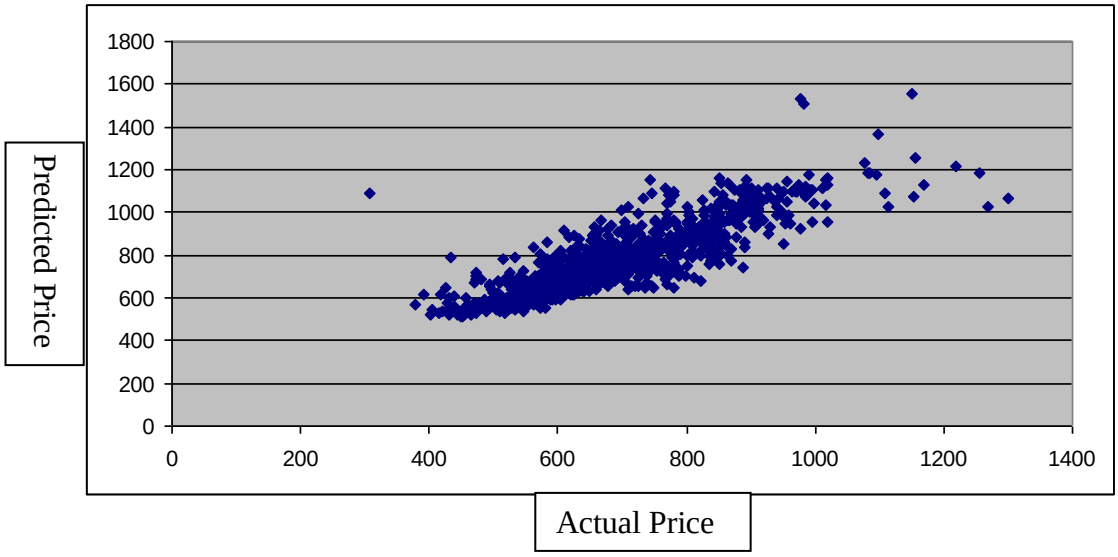


Figure 4.16: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of October 2001 with Bagging

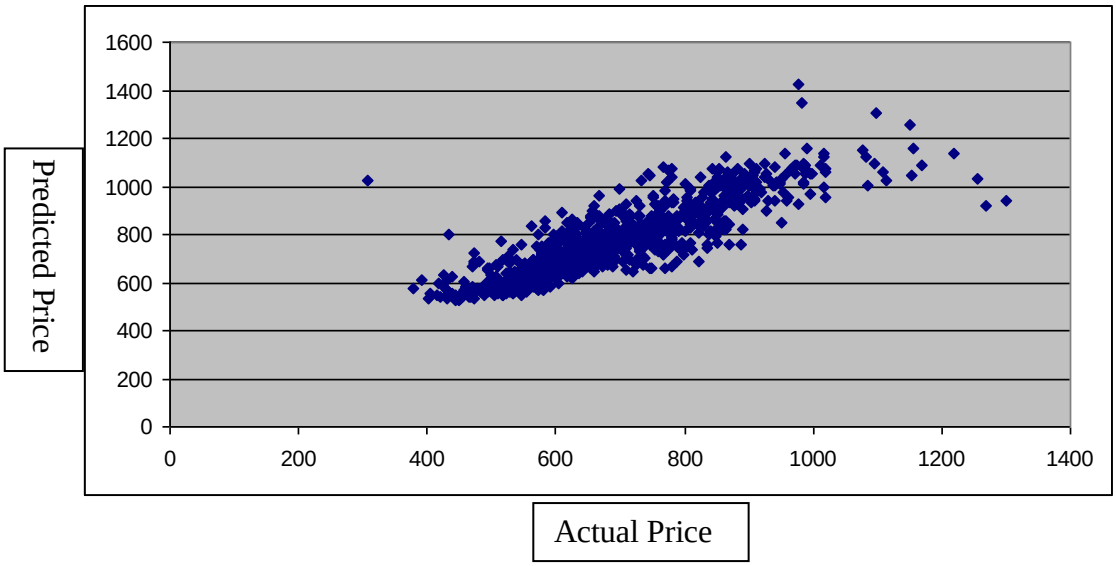


Figure 4.17: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of October 2001 with Random Forest

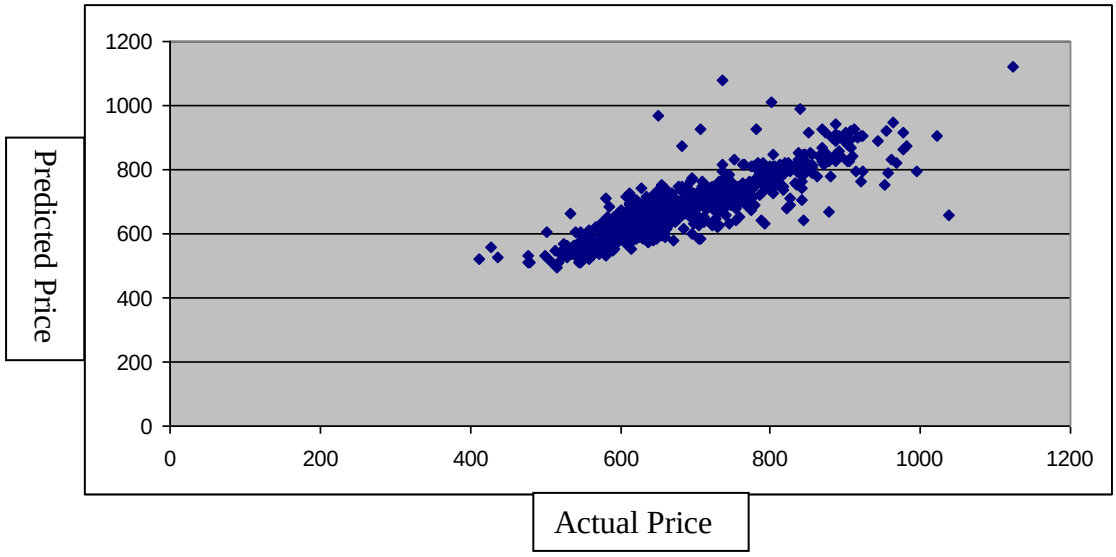


Figure 4.18: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of November 2001 with Bagging

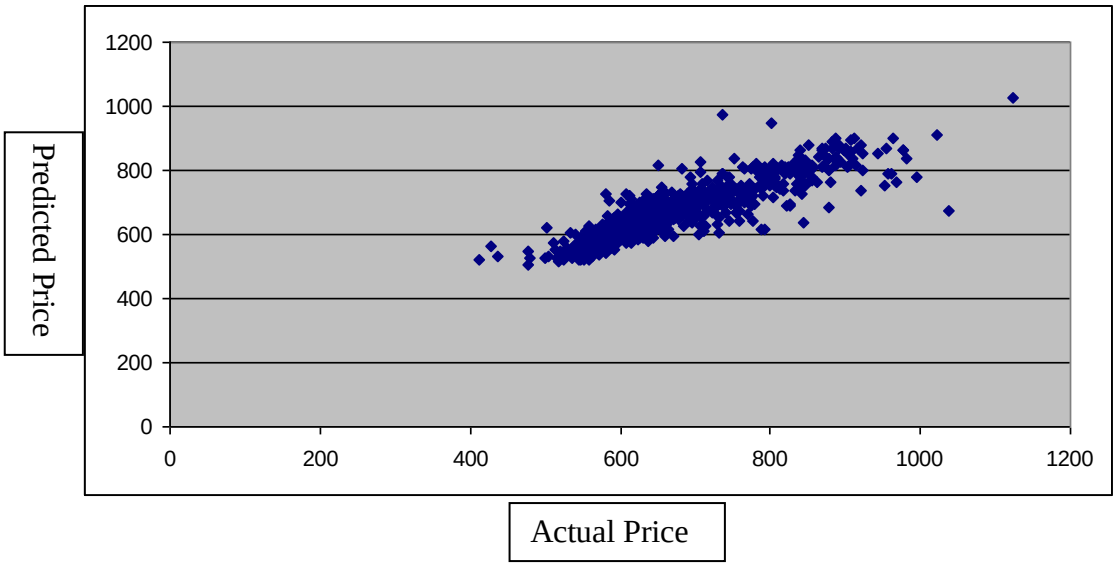


Figure 4.19: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of November 2001 with Random Forest

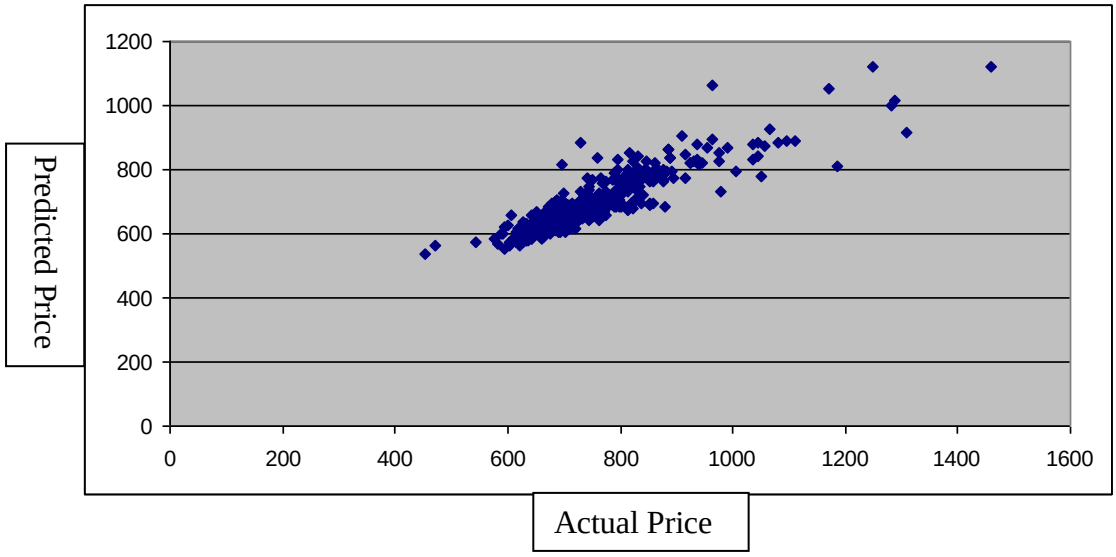


Figure 4.20: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of December 2001 with Bagging

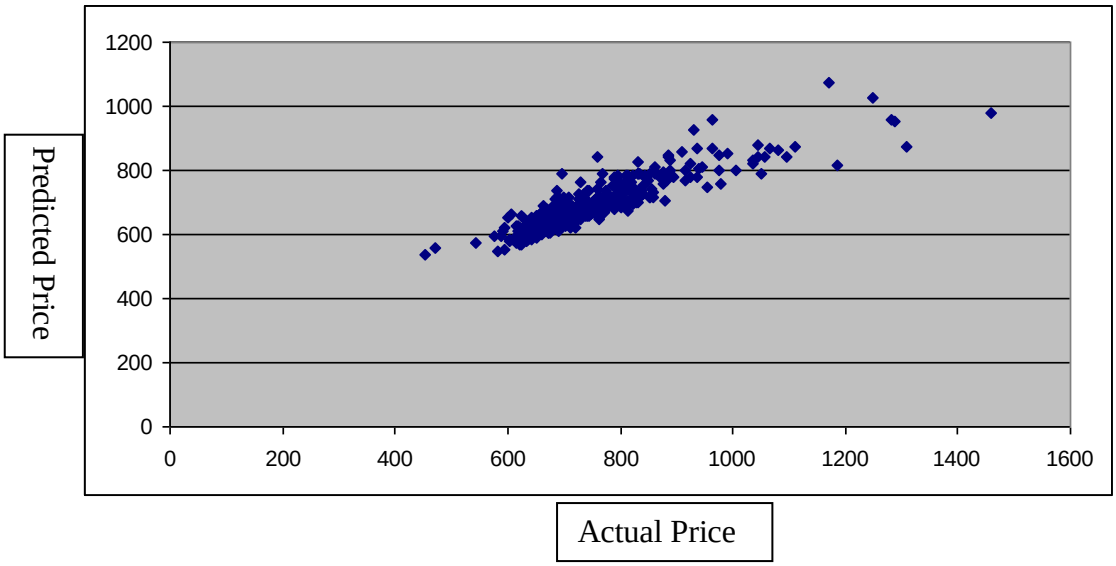


Figure 4.21: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of December 2001 with Random Forest

Table 4.22: Predictions for Period C with Ensemble Methods

| | | GRNN | Regression Tree (without pruning) | Bagging | Random Forest |
|---|---|---|---|---|---|
| Using last week of Aug 2002 to predict 1st wk of Sep 2002 | Root Mean Square Error | 38.60712508 | 49.9932 | 38.81197 | 38.20277 |
| | Mean Absolute Error | 27.09162225 | 32.91185 | 26.6451 | 26.55978 |
| | Std. Deviation of Abs. Error | 27.51594552 | 37.64576 | 28.23138 | 27.47001 |
| Using last week of Sep 2002 to predict 1st wk of Oct 2002 | Root Mean Square Error | 67.49464254 | 77.1083 | 66.68828 | 65.17472 |
| | Mean Absolute Error | 42.78633291 | 49.20164 | 42.83662 | 42.84812 |
| | Std. Deviation of Abs. Error | 52.20917584 | 59.38102 | 51.11999 | 49.11839 |
| Using last week of Oct 2002 to predict 1st wk of Nov 2002 | Root Mean Square Error | 53.65067084 | 60.417 | 53.29528 | 50.8611 |
| | Mean Absolute Error | 38.96359524 | 43.58813 | 39.29378 | 37.70949 |
| | Std. Deviation of Abs. Error | 36.89281363 | 41.84948 | 36.01656 | 34.14046 |
| Using last week of Nov 2002 to predict 1st wk of Dec 2002 | Root Mean Square Error | 41.68142591 | 46.99935 | 40.80626 | 41.99026 |
| | Mean Absolute Error | 28.36965042 | 32.1573 | 28.09967 | 28.62148 |
| | Std. Deviation of Abs. Error | 30.54913432 | 34.28974 | 29.60168 | 30.73675 |

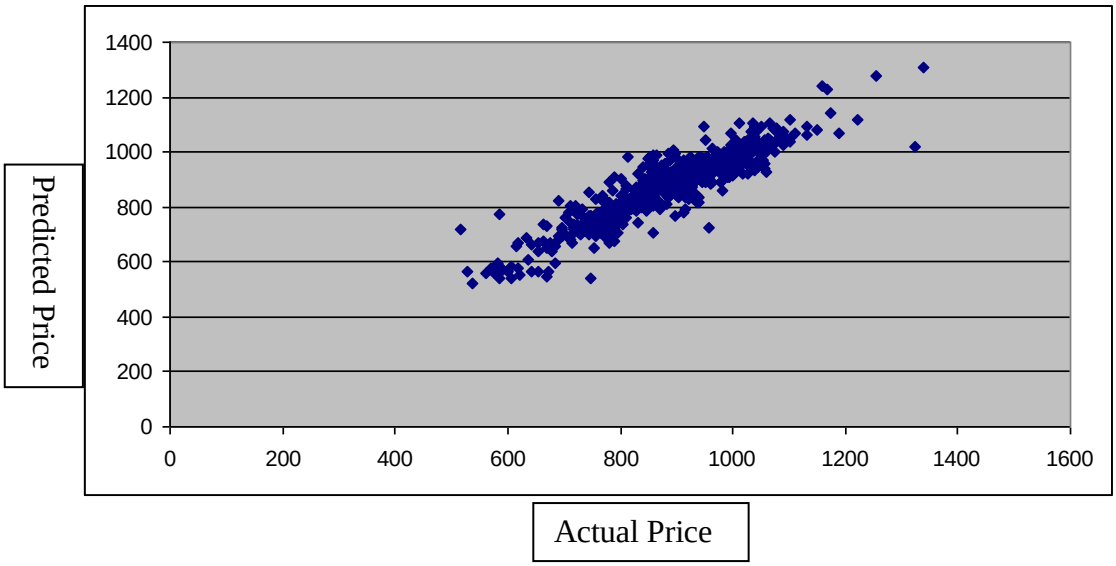


Figure 4.22: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of September 2002 with Bagging

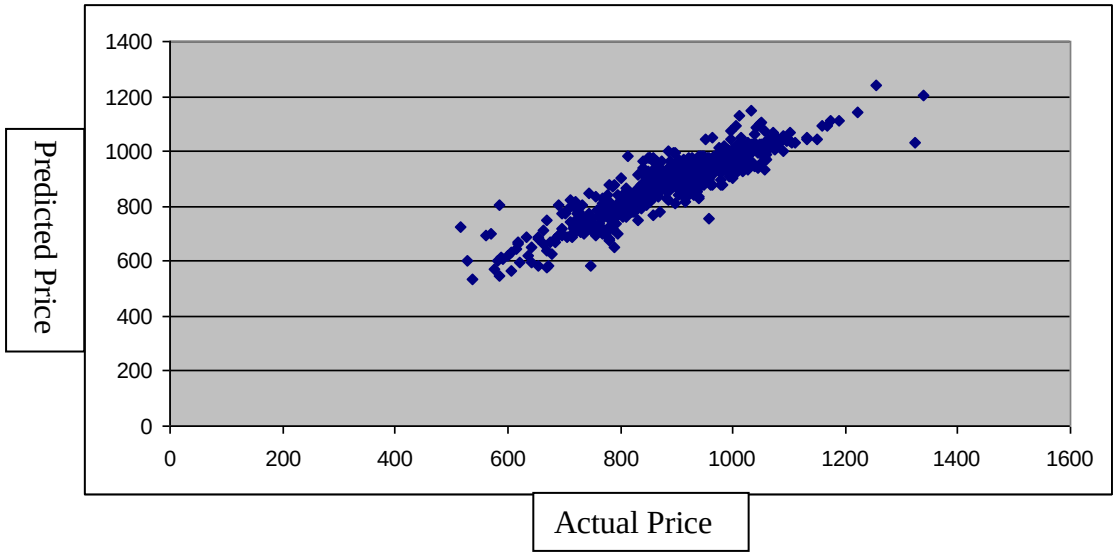


Figure 4.23: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of September 2002 with Random Forest

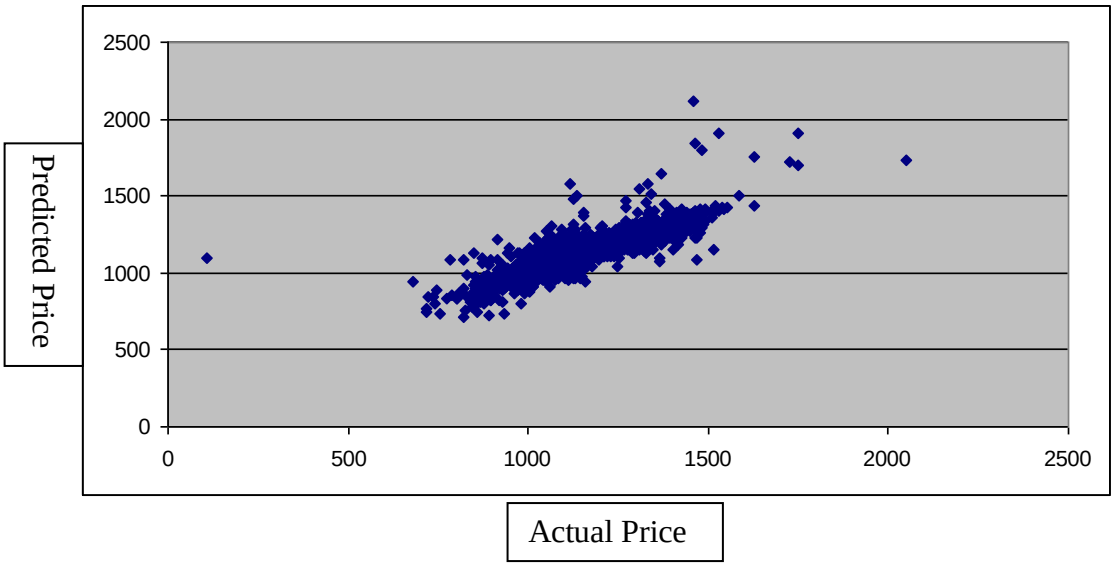


Figure 4.24: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of October 2002 with Bagging

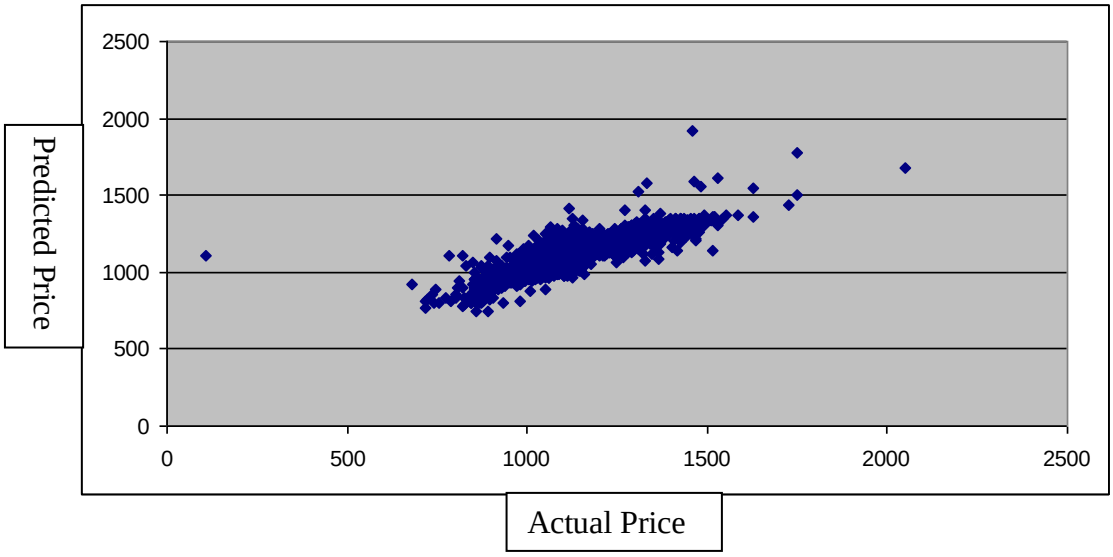


Figure 4.25: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of October 2002 with Random Forest

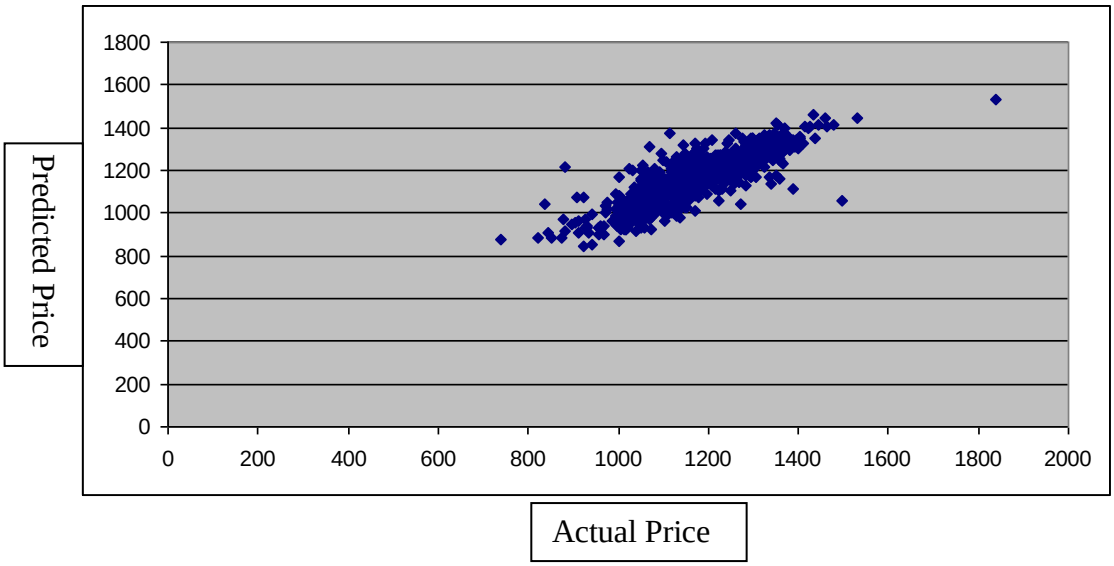


Figure 4.26: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of November 2002 with Bagging

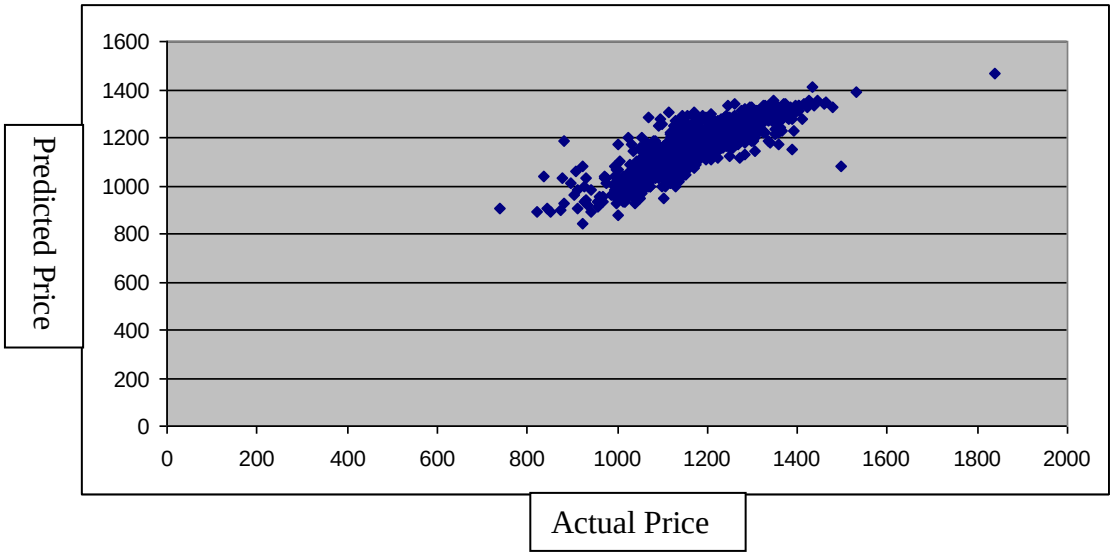


Figure 4.27: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of November 2002 with Random Forest

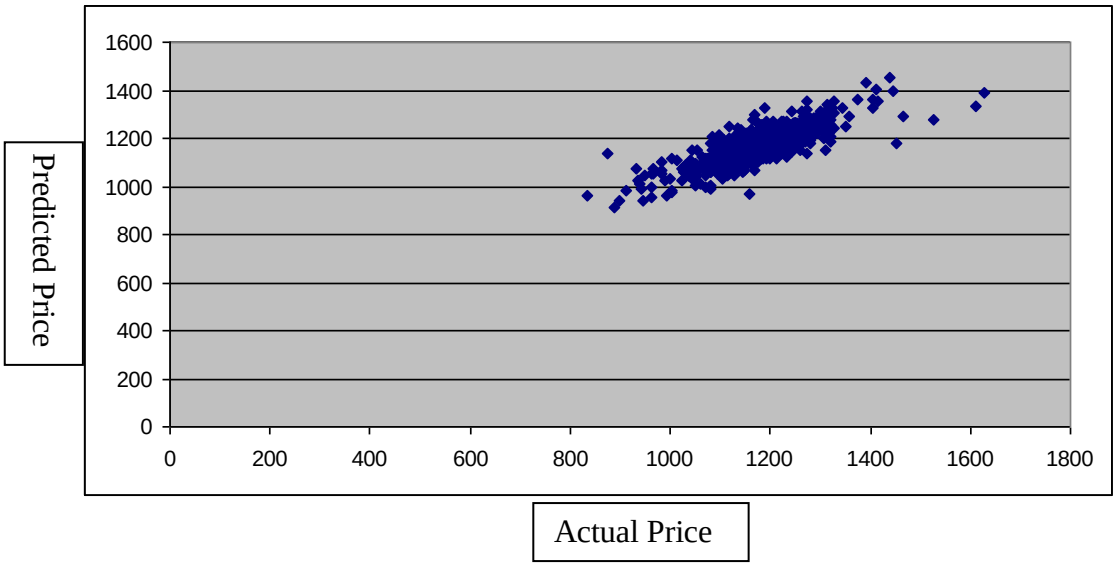


Figure 4.28: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of December 2002 with Bagging

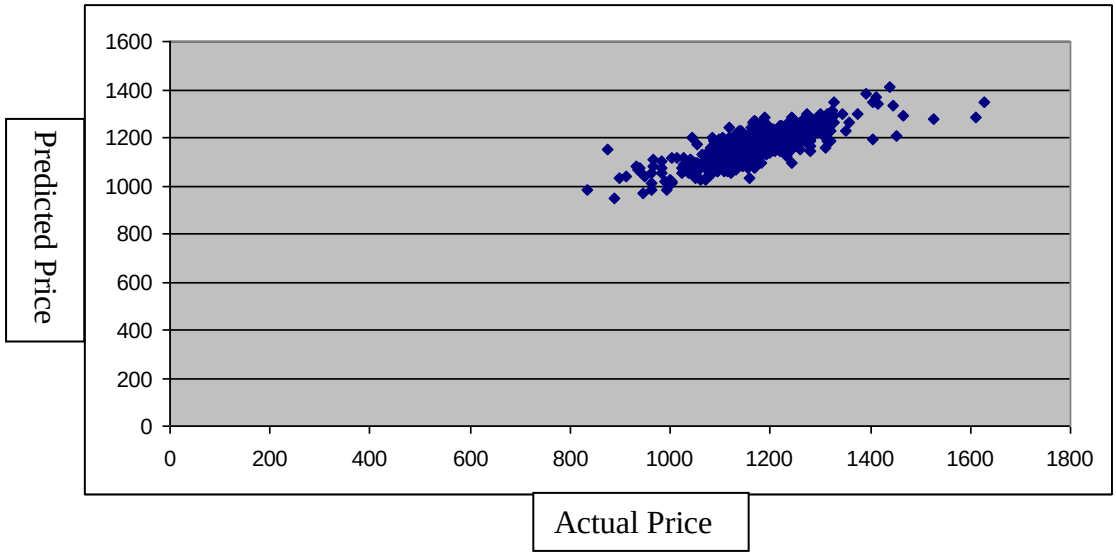


Figure 4.29: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of December 2002 with Random Forest

## 4.7 Discussions on Tree-based Regression Methods

In Sections 4.1 we have shown the regression tree's advantage over methods such as neural networks. The trees generated are able to detect interaction between parts of levels or parts of the numeric range of independent variables.

They can be interpreted and give us better understanding of the price-driving variables, satisfying the descriptive aspect of modelling. It is possible to tell the order of importance of the wool quality variables and their influences in driving the price at various levels.

Section 4.4 showed another huge advantage of the tree models: their ability to provide helpful information to solve the wool specifications problem introduced in Section 1.3. This huge advantage is exclusive to the tree models and not available from neural networks or other existing methods.

However, from the results presented in Section 4.5 we can observe that regression tree's accuracies in both fitting and predictions are quite poor when compared to methods such as neural networks. To rectify this we introduce the ensemble methods in Section 4.6 and apply them to the wool auction data. We found that by averaging the numeric price outputs from multiple trees we can produce much better results with accuracies on par with those from neural networks.

However, while a prediction is numeric and can be obtained by averaging the output from each tree, the actual trees themselves cannot be "added" and "averaged" numerically. Here, we lose the simplicity of having only one tree, and the advantages hence attractiveness that comes with them. So, finding a single tree that is the equivalent to an "average" of multiple trees becomes an interesting problem. And we shall explore this concept in the next chapter.

# Chapter 5

# A Hybrid Approach

In the previous chapter, we learnt that we could improve the prediction results from tree predictors by averaging the outputs from multiple predictors, all constructed using the same data. In the application of bagging, we can take multiple bootstrap samples from the learning set of data, then grow a tree from each bootstrap sample. The predictions from these trees can then be averaged to give us an overall prediction, which is a much improved prediction than a single ordinary regression tree. In essence, the bootstrap procedure introduced randomness which reduced variance in the data, so the model built from it can be more accurate. There exist variants of the bagging method such as random forests which also yield similar results. However, while a prediction is numeric and can be obtained by averaging the output from each tree, the actual trees themselves cannot be "added" and "averaged" numerically. Here, we lose the simplicity of having only one tree, and the advantages hence attractiveness that comes with them. So, finding a single tree that is the equivalent to an "average" of multiple trees becomes an interesting problem.

Attempting to provide a solution to the above problem, Breiman and Shang (1997) introduced the idea of a "representer tree" (also known as a "born again tree"). Unlike multiple tree methods where you make a specified number (say 10) of bootstrap samples of the same dataset then use these to build 10 trees, their idea involved manufacturing an artificial dataset that is 10 times the size of the original and used the new set to build a single tree. Their study showed that this alternate approach gave prediction accuracy comparable to those of the multiple tree models.

As a result of our work in Chapter 3 and Chapter 4, we will develop and apply a new procedure to the wool auction problem in this chapter. Our procedure is inspired by Breiman and Shang's idea but with two main modifications: 1) using

a much more accurate Neural Network in place of a multiple tree method, and 2) using our own modified smearing method which involves adding Gaussian noise.

This chapter concludes that our new hybrid approach is the best balance between prediction accuracies and interpretability that we can currently achieve. It provides solutions to the various aspects of our initial wool auction problem, as well as the wool specifications problem discussed in Section 1.3.

## 5.1 The Method

Keeping the ideas and terminologies from Chapter 2 in mind, let us consider the following data set:

Table 5.1: Example Data Set

| | Variable $x_1$ | Variable $x_2$ | Variable $x_3$ | … | Variable $x_n$ | Price $y$ |
|---|---|---|---|---|---|---|
| Data 1 | $x_{11}$ | $x_{12}$ | $x_{13}$ | | $X_{1n}$ | $y_1$ |
| Data 2 | $x_{21}$ | $x_{22}$ | $x_{23}$ | | $X_{2n}$ | $y_2$ |
| Data 3 | $x_{31}$ | $x_{32}$ | $x_{33}$ | … | $X_{3n}$ | $y_3$ |
| Data 4 | $x_{41}$ | $x_{42}$ | $x_{43}$ | | $X_{4n}$ | $y_4$ |
| Data 5 | $x_{51}$ | $x_{52}$ | $x_{53}$ | | $X_{5n}$ | $y_5$ |
| . . . | | | | | | |

Say we have built a predictive model $\tilde{f}$ from the data. Breiman and Shang (1997) considered multiple tree predictors such as bagging, but here we will use neural networks such as GRNN instead. While both bagging and GRNN offered similar prediction accuracy, we observed in Section 4.5 that bagging is relatively poor in fitting while GRNN give consistently good results. As a result we have decided to use GRNN in our approach instead.

To grow a single regression tree from the data, we need to generate a new set of artificial data that is much bigger in size than the original, but with the same underlying multivariate distribution of the original. The problem of manufacturing a large amount of artificial data was earlier considered by Craven

and Shavlik (1996). Their approach involved constructing a kernel density estimate for each input variable separately, then sampled from the product of the density estimates.

Later, Breiman and Shang (1997) grew a much bigger data set by smearing the original data, which gave them better results. First they randomly picked a row of data. For each variable in this row, a random number was generated that was between 0 and 1. If this random number was greater than 0.5 (the palt) then they did not change the value of the variable. If the random number was less than 0.5 then the value of the variable was replaced with one from the same column but in another random row, hence "smearing" the data. This was done for every variable in each row. And they generated as many new rows this way as needed. The Price $y$ could then be given by a multiple tree predictor $f$. A typical example of smearing is displayed in Table 5.2 below. Say the original data set only has 5 rows, and we take the first 5 rows of data from Table 5.1 as a demonstration. If we want to generate an artificial data set that is 10 times as big then we would pick a random row from the original 5 rows and do this 50 times.

Table 5.2: Example of 50% Smearing.

| | Variable $x_1$ | Variable $x_2$ | Variable $x_3$ | … | Variable $x_n$ | Price $y$ |
|---|---|---|---|---|---|---|
| Row 3 | $x_{31}$ (no change) | $x_{52}$ (smeared) | $x_{43}$ (smeared) | | | $\tilde{f}(x_{31}, x_{52}, x_{43}, \ldots)$ |
| Row 5 | $x_{11}$ (smeared) | $x_{42}$ (smeared) | $x_{53}$ (no change) | | | $\tilde{f}(x_{11}, x_{42}, x_{53}, \ldots)$ |
| Row 1 | $x_{11}$ (no change) | $x_{32}$ (smeared) | $x_{13}$ (no change) | | | $f(x_{11}, x_{32}, x_{13}, \ldots)$ |
| Row 3 | $x_{41}$ (smeared) | $x_{32}$ (no change) | $x_{33}$ (no change) | | | $f(x_{41}, x_{32}, x_{33}, \ldots)$ |
| Row 2 | | | | | | . |
| Row 5 | | | | | | . |
| Row 4 | | | | | | . |
| Row 4 | | | | … | … | |
| . . . | | | | | | |

We could also use a number other than 0.5 as our palt number. Breiman and Shang (1997) called this a 0.5 palt a "50% smearing"; a 0.25 palt is "25% smearing".

The above smearing method was designed allowing any variable to be numeric or categorical. However, since the wool data we are using consists of purely

numeric variables, we are free to make the following modifications when applying the method to our wool data. When we smear the value of a variable, instead of replacing it with one from another row, we keep the old value but add Gaussian noise to it. So

$$\text{smeared value} = \text{original value of the variable} + p * N(0, \sigma^2) \qquad (5.1)$$

Here $\sigma^2$ is the variance of the variable (the column) and "$p$" is a parameter we can set to determine how far we want to smear the value. We will demonstrate in the next section that our modification with a $p$ value of about 0.5 in makes the underlying multivariate distribution even closer to that of the original data than Breiman and Shang's original 50% smearing. The tree grown from the artificial data smeared this way also gives better results in predictions.

The underlying multivariate distribution of the artificial set very closely mimics the distribution of the original data set. Of course, this also depends on how good $f$ (in our case the neural network) is in approximating the real $f$.

Using the variables Staple Length and Staple Strength from a particular day of our wool data as an illustrative example, Figures 5.1, 5.2 and 5.3 demonstrate the effects and differences from the two smearing methods. The figures clearly show that our modified smearing method can retain the "shape" of the distribution between two wool variables, making our artificially manufactured data a more realistic representation of the original data.

Once we have this much bigger set of data then we can grow a single regression tree from it. This "representer tree" would be much more accurate than one grown directly from the original data set.

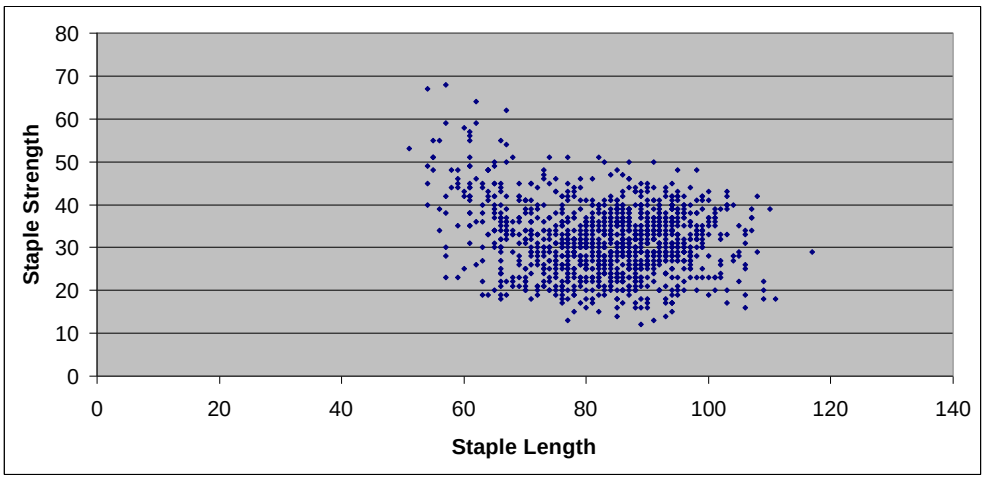


Figure 5.1: Original Data

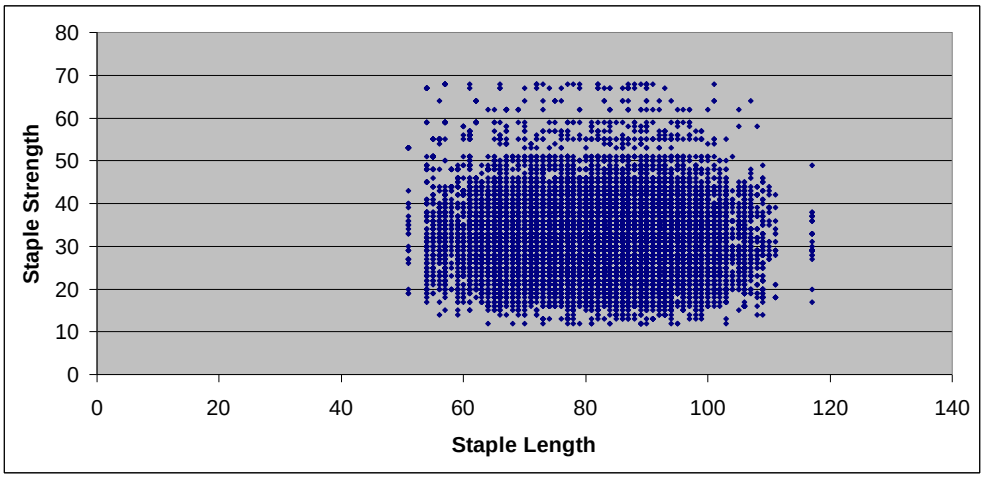


Figure 5.2: 50% Smearing (Note that this smearing mimics the proportion of density distribution, but does not retain the shape of original data.)

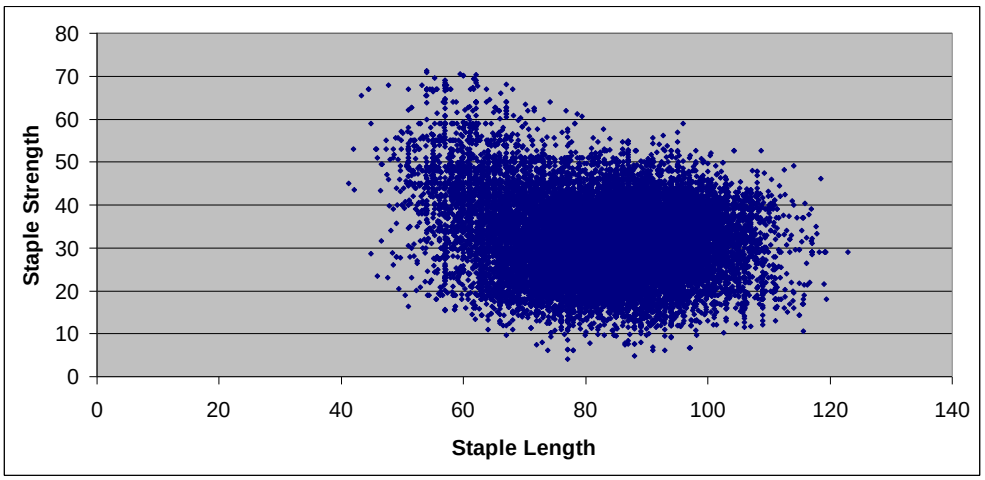


Figure 5.3: An example of our Modified Smearing using Gaussian Noise (Here, the modified method preserves the proportion of density distribution as well as the shape.)

## 5.2 Comparisons of Algorithms: Neural Networks vs. Regression Tree vs. Ensemble Methods vs. Hybrid Approach

In this section we present diagrammatic summaries of the algorithms used in regression tree, bagging, and our hybrid approach.

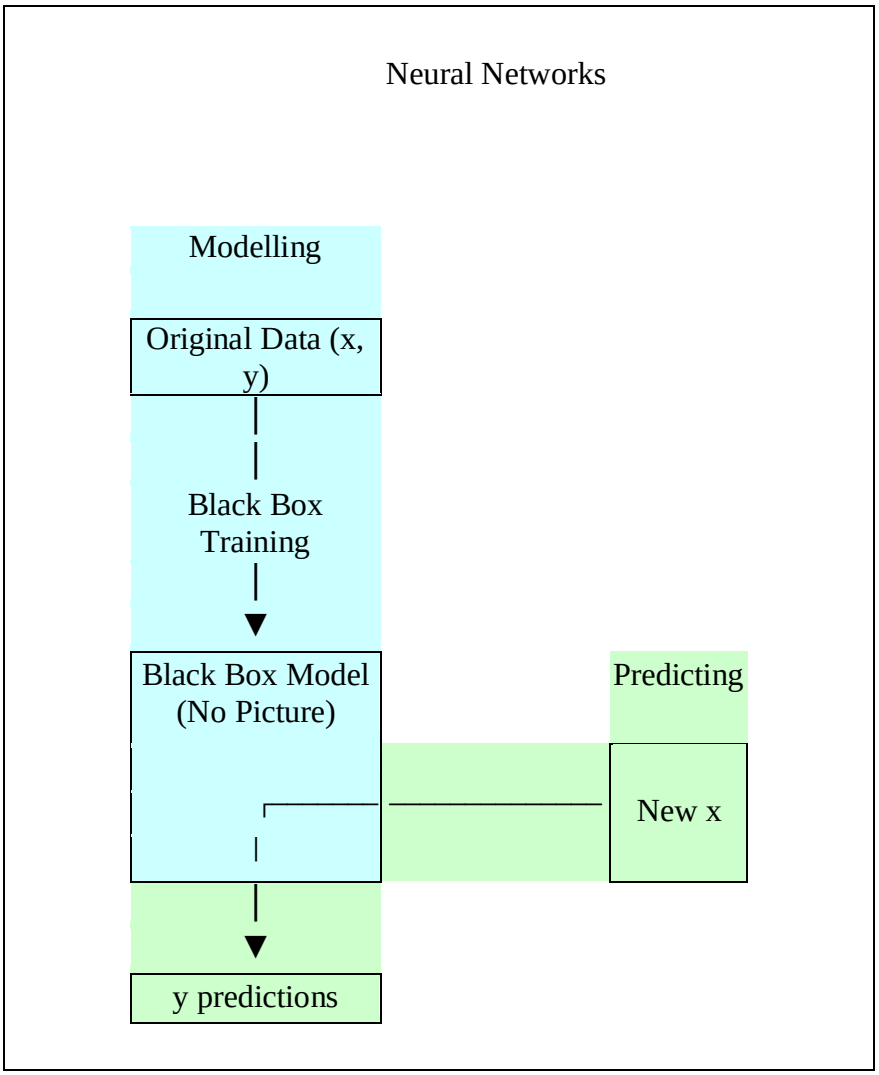


Figure 5.4: Diagrammatic Summary of Neural Networks

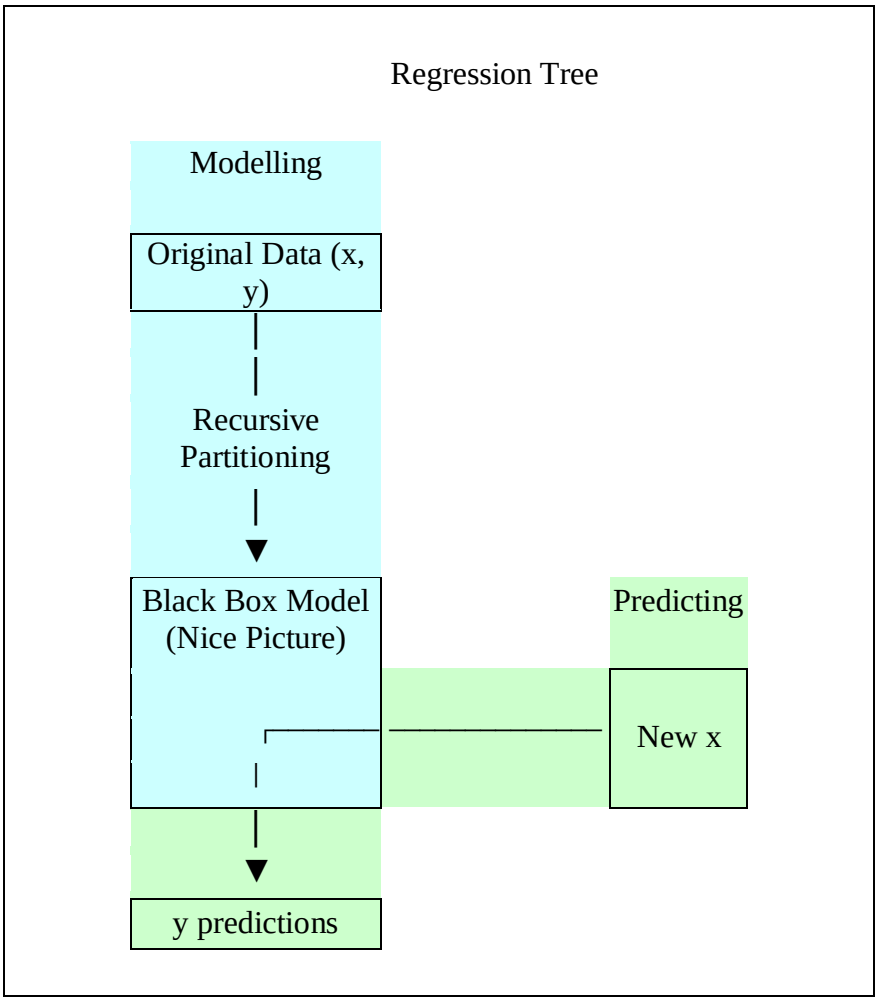


Figure 5.5: Diagrammatic Summary of Regression Tree

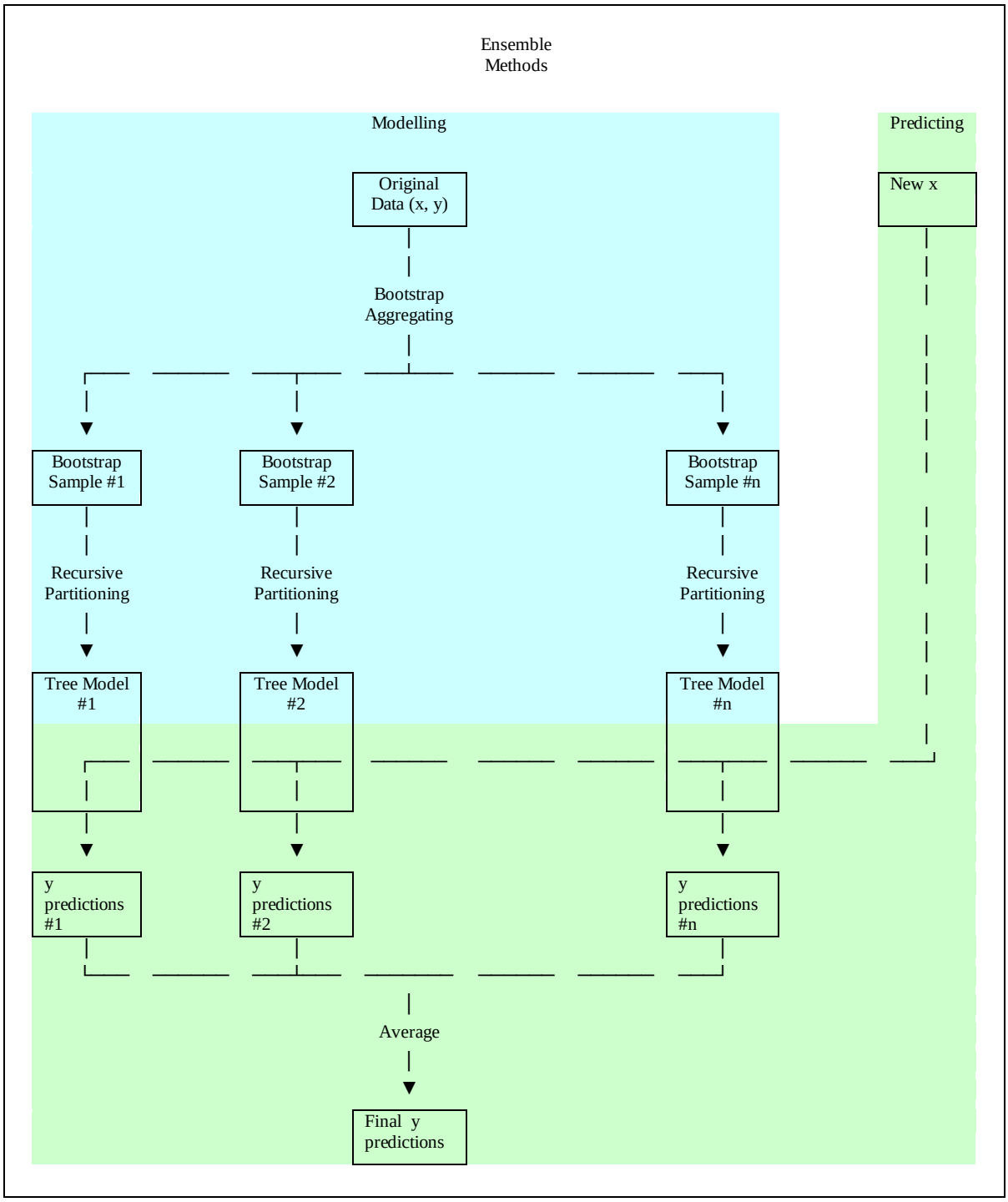


Figure 5.6: Diagrammatic Summary of Ensemble Methods

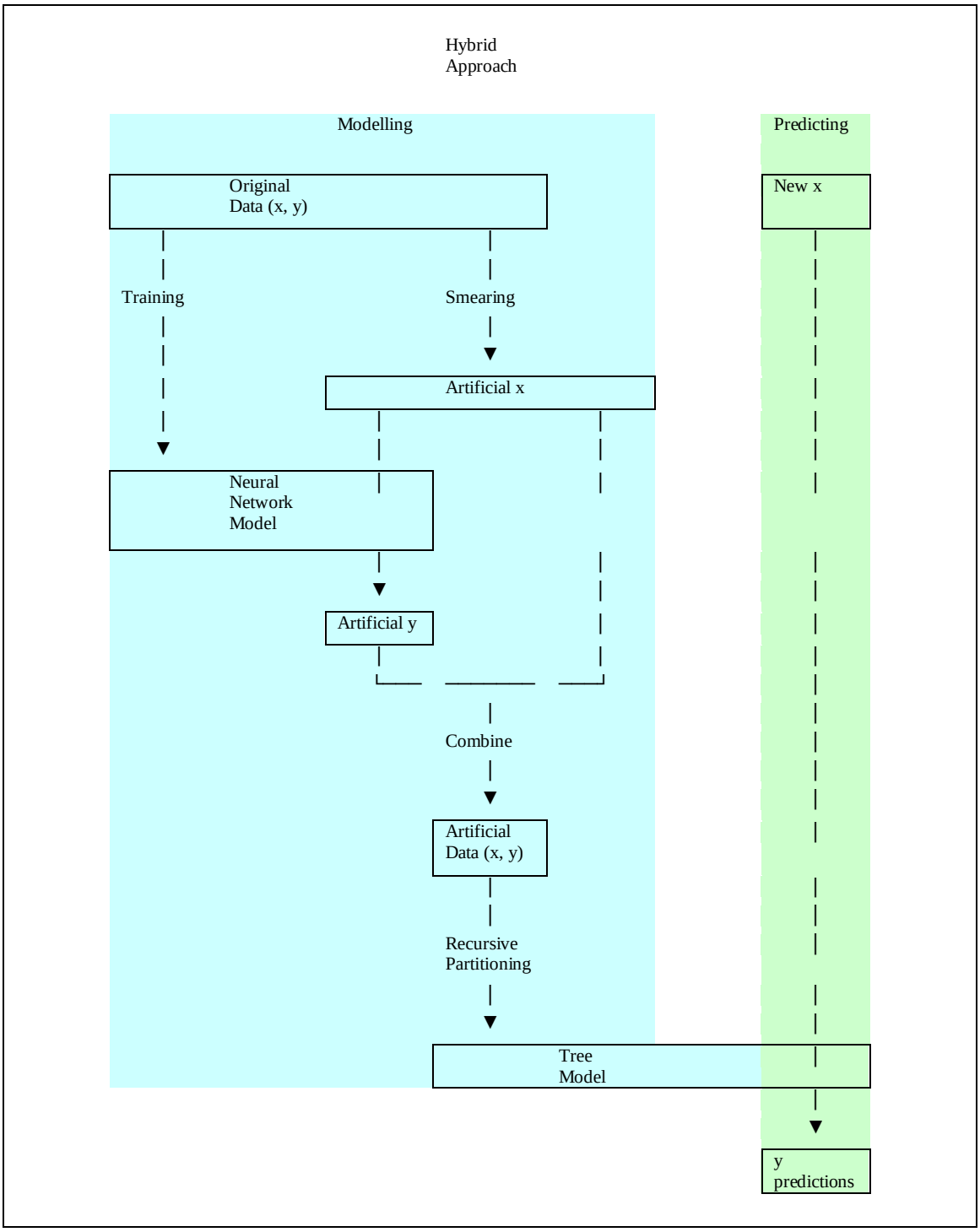


Figure 5.7: Diagrammatic Summary of Hybrid Approach

## 5.3 Applying Our Hybrid Approach to the Wool Auction Data

Before comparing the hybrid approach to the earlier methods such as GRNN, regression tree, and ensemble methods, we will first determine a reasonable *p* value to be used in our modified smearing.

The tables that follow show the results from representer trees (with or without pruning as explained in Section 4.3) generated with Breiman and Shang's original 50% smearing, as well as our own modified smearing with some different *p* values. From these tables, we find our modified smearing with a *p* value of 0.5 (and without pruning the tree) gives the best results in both fitting and predictions. We shall use a hybrid approach with this particular smearing for comparison with the earlier modelling methods.

Table 5.3: Fitting for Period A with Hybrid Approach

| | | Hybrid Approach | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | (50% smearing, not pruned) | (50% smearing, pruned) | (modified smearing with p = 0.5, not pruned) | (modified smearing with p = 0.5, pruned) | (modified smearing with p = 1, not pruned) | (modified smearing with p = 1, pruned) | (modified smearing with p = 2, not pruned) | (modified smearing with p = 2, pruned) |
| Fitting last week of Aug 2000 | Root Mean Square Error | 32.14784 | 32.43423 | 28.85759 | 29.51178 | 29.46676 | 30.15601 | 30.81464 | 32.60011 |
| | Mean Absolute Error | 18.05 | 18.70416 | 15.60698 | 16.56492 | 16.29405 | 17.344 | 17.05616 | 19.04959 |
| | Std. Deviation of Abs. Error | 26.61113 | 26.50662 | 24.28116 | 24.43248 | 24.56004 | 24.67743 | 25.67232 | 26.46406 |
| Fitting last week of Sep 2000 | Root Mean Square Error | 40.52424 | 41.09794 | 33.95739 | 34.81306 | 35.64119 | 36.27338 | 37.63739 | 39.78697 |
| | Mean Absolute Error | 20.21789 | 21.29925 | 16.98274 | 18.2082 | 18.18383 | 19.15071 | 19.35314 | 21.74229 |
| | Std. Deviation of Abs. Error | 35.12982 | 35.15731 | 29.41342 | 29.67957 | 30.66171 | 30.81413 | 32.28902 | 33.32962 |
| Fitting last week of Oct 2000 | Root Mean Square Error | 41.70309 | 42.68297 | 36.31789 | 37.40157 | 37.0496 | 38.71816 | 39.33116 | 43.05612 |
| | Mean Absolute Error | 23.84127 | 25.00416 | 21.23019 | 22.68594 | 21.78662 | 23.47373 | 22.66546 | 26.71006 |
| | Std. Deviation of Abs. Error | 34.22557 | 34.60189 | 29.47456 | 29.74417 | 29.97521 | 30.79943 | 32.15259 | 33.77919 |
| Fitting last week of Nov 2000 | Root Mean Square Error | 53.91381 | 54.61658 | 53.43675 | 53.69071 | 53.72859 | 54.26177 | 54.62545 | 55.87569 |
| | Mean Absolute Error | 19.72541 | 21.43055 | 17.88617 | 18.68215 | 18.38907 | 19.94024 | 19.44443 | 22.48246 |
| | Std. Deviation of Abs. Error | 50.1974 | 50.25813 | 50.37617 | 50.35729 | 50.50547 | 50.48687 | 51.06958 | 51.17508 |

Table 5.4: Fitting for Period B with Hybrid Approach

| | | Hybrid Approach | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | (50% smearing, not pruned) | (50% smearing, pruned) | (modified smearing with p = 0.5, not pruned) | (modified smearing with p = 0.5, pruned) | (modified smearing with p = 1, not pruned) | (modified smearing with p = 1, pruned) | (modified smearing with p = 2, not pruned) | (modified smearing with p = 2, pruned) |
| Fitting last week of Aug 2001 | Root Mean Square Error | 40.0757 | 40.65163 | 36.11205 | 36.45528 | 37.23485 | 37.59594 | 38.58073 | 39.83728 |
| | Mean Absolute Error | 22.1232 | 22.80188 | 19.56654 | 20.31361 | 20.17789 | 21.10055 | 21.35856 | 22.99529 |
| | Std. Deviation of Abs. Error | 33.42948 | 33.66818 | 30.36407 | 30.28343 | 31.30622 | 31.12885 | 32.14218 | 32.54354 |
| Fitting last week of Sep 2001 | Root Mean Square Error | 48.2608 | 48.79781 | 44.28918 | 44.70897 | 46.90276 | 47.48962 | 46.36615 | 47.38853 |
| | Mean Absolute Error | 22.23308 | 23.37491 | 19.81226 | 20.89538 | 21.067 | 22.33507 | 21.79332 | 23.51032 |
| | Std. Deviation of Abs. Error | 42.84418 | 42.8447 | 39.61962 | 39.53455 | 41.91472 | 41.91899 | 40.93442 | 41.15461 |
| Fitting last week of Oct 2001 | Root Mean Square Error | 32.683 | 33.01007 | 29.78828 | 30.03139 | 30.35766 | 30.67092 | 31.81951 | 32.82703 |
| | Mean Absolute Error | 18.39031 | 19.04518 | 16.45713 | 16.89795 | 17.10047 | 17.87151 | 17.48543 | 18.80197 |
| | Std. Deviation of Abs. Error | 27.03093 | 26.97478 | 24.84133 | 24.83811 | 25.09504 | 24.93806 | 26.59727 | 26.92192 |
| Fitting last week of Nov 2001 | Root Mean Square Error | 22.37742 | 23.18776 | 21.57403 | 21.7966 | 21.66333 | 22.4787 | 22.87607 | 24.16405 |
| | Mean Absolute Error | 12.64958 | 13.98992 | 11.51276 | 12.18102 | 11.73037 | 13.05542 | 12.6078 | 14.58006 |
| | Std. Deviation of Abs. Error | 18.46909 | 18.50204 | 18.25531 | 18.08505 | 18.22247 | 18.30878 | 19.09853 | 19.28021 |

Table 5.5: Fitting for Period C with Hybrid Approach

| | | Hybrid Approach | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | (50% smearing, not pruned) | (50% smearing, pruned) | (modified smearing with p = 0.5, not pruned) | (modified smearing with p = 0.5, pruned) | (modified smearing with p = 1, not pruned) | (modified smearing with p = 1, pruned) | (modified smearing with p = 2, not pruned) | (modified smearing with p = 2, pruned) |
| Fitting last week of Aug 2002 | Root Mean Square Error | 32.67029 | 33.38364 | 28.32042 | 28.46928 | 29.73316 | 30.90868 | 31.1653 | 33.09808 |
| | Mean Absolute Error | 18.60128 | 19.59546 | 16.03741 | 16.79665 | 16.89491 | 18.43153 | 18.1521 | 20.15263 |
| | Std. Deviation of Abs. Error | 26.86778 | 27.03756 | 23.35065 | 22.99491 | 24.47587 | 24.82103 | 25.34275 | 26.26533 |
| Fitting last week of Sep 2002 | Root Mean Square Error | 43.50797 | 44.63964 | 40.10445 | 40.9717 | 41.50845 | 43.32336 | 44.16638 | 47.67187 |
| | Mean Absolute Error | 29.42514 | 30.72236 | 26.16641 | 27.54871 | 27.30787 | 29.6761 | 29.08741 | 32.90743 |
| | Std. Deviation of Abs. Error | 32.05945 | 32.39679 | 30.40262 | 30.33775 | 31.27141 | 31.57411 | 33.24678 | 34.50395 |
| Fitting last week of Oct 2002 | Root Mean Square Error | 35.79199 | 36.16242 | 33.79716 | 34.1464 | 35.54239 | 35.7953 | 37.40348 | 38.44765 |
| | Mean Absolute Error | 22.32846 | 22.84543 | 19.86391 | 20.44141 | 20.85458 | 21.5572 | 22.33803 | 23.79862 |
| | Std. Deviation of Abs. Error | 27.98414 | 28.04309 | 27.35419 | 27.36247 | 28.79217 | 28.58711 | 30.01215 | 30.2085 |
| Fitting last week of Nov 2002 | Root Mean Square Error | 34.39373 | 34.64131 | 33.72553 | 33.95238 | 34.26232 | 34.54748 | 34.6122 | 34.78557 |
| | Mean Absolute Error | 22.40636 | 22.57759 | 21.79123 | 22.0425 | 21.98507 | 22.28514 | 22.33603 | 22.6841 |
| | Std. Deviation of Abs. Error | 26.10821 | 26.2876 | 25.75438 | 25.83857 | 26.29313 | 26.41351 | 26.45527 | 26.38634 |

Table 5.6: Predictions for Period A with Hybrid Approach

| | | Hybrid Approach | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | (50% smearing, not pruned) | (50% smearing, pruned) | (modified smearing with p = 0.5, not pruned) | (modified smearing with p = 0.5, pruned) | (modified smearing with p = 1, not pruned) | (modified smearing with p = 1, pruned) | (modified smearing with p = 2, not pruned) | (modified smearing with p = 2, pruned) |
| Using last week of Aug 2000 to predict 1st wk of Sep 2000 | Root Mean Square Error | 51.25978 | 51.29458 | 51.4169 | 51.54048 | 51.44226 | 51.60238 | 52.15323 | 52.65016 |
| | Mean Absolute Error | 29.59563 | 29.66381 | 29.04302 | 29.18406 | 29.27836 | 29.65777 | 30.00095 | 30.42627 |
| | Std. Deviation of Abs. Error | 41.86422 | 41.85858 | 42.44027 | 42.49341 | 42.30902 | 42.23965 | 42.67186 | 42.98001 |
| Using last week of Sep 2000 to predict 1st wk of Oct 2000 | Root Mean Square Error | 136.3895 | 136.5583 | 129.6546 | 129.7598 | 137.4913 | 137.2228 | 138.683 | 138.8697 |
| | Mean Absolute Error | 32.14405 | 32.39875 | 29.77496 | 30.16947 | 31.72107 | 31.95903 | 32.19726 | 33.46321 |
| | Std. Deviation of Abs. Error | 132.5985 | 132.7103 | 126.2379 | 126.2524 | 133.8334 | 133.5005 | 134.9455 | 134.8294 |
| Using last week of Oct 2000 to predict 1st wk of Nov 2000 | Root Mean Square Error | 78.87132 | 79.48225 | 74.0424 | 74.8116 | 74.80524 | 75.22768 | 75.5165 | 76.69695 |
| | Mean Absolute Error | 39.77801 | 39.9867 | 37.49623 | 38.14071 | 38.46077 | 38.39608 | 38.14695 | 39.94294 |
| | Std. Deviation of Abs. Error | 68.12156 | 68.70721 | 63.86079 | 64.37378 | 64.17556 | 64.70615 | 65.18836 | 65.49024 |
| Using last week of Nov 2000 to predict 1st wk of Dec 2000 | Root Mean Square Error | 63.68896 | 63.68942 | 61.89901 | 62.19532 | 67.32929 | 67.59487 | 69.49303 | 70.11804 |
| | Mean Absolute Error | 35.32574 | 35.53997 | 35.34225 | 35.70845 | 35.59939 | 36.12464 | 36.31719 | 37.01193 |
| | Std. Deviation of Abs. Error | 53.01956 | 52.87661 | 50.84185 | 50.94757 | 57.17565 | 57.15953 | 59.2766 | 59.5824 |

Table 5.7: Predictions for Period B with Hybrid Approach

| | | Hybrid Approach | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | (50% smearing, not pruned) | (50% smearing, pruned) | (modified smearing with p = 0.5, not pruned) | (modified smearing with p = 0.5, pruned) | (modified smearing with p = 1, not pruned) | (modified smearing with p = 1, pruned) | (modified smearing with p = 2, not pruned) | (modified smearing with p = 2, pruned) |
| Using last week of Aug 2001 to predict 1st wk of Sep 2001 | Root Mean Square Error | 62.76642 | 62.90683 | 60.49215 | 60.55243 | 61.75273 | 61.86352 | 64.29749 | 64.34594 |
| | Mean Absolute Error | 32.98242 | 33.08886 | 32.53699 | 32.63289 | 33.21945 | 33.41471 | 33.73901 | 33.87366 |
| | Std. Deviation of Abs. Error | 53.4211 | 53.52041 | 51.01466 | 51.02493 | 52.07492 | 52.08149 | 54.7538 | 54.72755 |
| Using last week of Sep 2001 to predict 1st wk of Oct 2001 | Root Mean Square Error | 90.3927 | 90.78546 | 88.2915 | 88.31937 | 88.71597 | 88.70276 | 90.1795 | 90.1926 |
| | Mean Absolute Error | 75.65191 | 75.99116 | 74.36393 | 74.39752 | 74.61819 | 74.63262 | 75.6311 | 75.43891 |
| | Std. Deviation of Abs. Error | 49.48291 | 49.68181 | 47.60521 | 47.60442 | 47.99505 | 47.94814 | 49.12427 | 49.44294 |
| Using last week of Oct 2001 to predict 1st wk of Nov 2001 | Root Mean Square Error | 39.64384 | 39.69916 | 38.57505 | 38.63218 | 39.7991 | 39.97535 | 39.81026 | 39.95935 |
| | Mean Absolute Error | 28.17994 | 28.23021 | 27.5051 | 27.6752 | 27.76045 | 27.81856 | 28.24166 | 28.14621 |
| | Std. Deviation of Abs. Error | 27.89296 | 27.92081 | 27.05489 | 26.96273 | 28.52789 | 28.71721 | 28.06714 | 28.37341 |
| Using last week of Nov 2001 to predict 1st wk of Dec 2001 | Root Mean Square Error | 67.54468 | 67.53837 | 67.10482 | 67.31418 | 68.69376 | 68.6115 | 66.99243 | 67.12219 |
| | Mean Absolute Error | 55.169 | 55.14255 | 54.33174 | 54.48521 | 55.41919 | 55.36777 | 55.03672 | 55.03495 |
| | Std. Deviation of Abs. Error | 38.99265 | 39.01916 | 39.40709 | 39.55215 | 40.6135 | 40.54445 | 38.2183 | 38.44808 |

Table 5.8: Predictions for Period C with Hybrid Approach

| | | Hybrid Approach | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | (50% smearing, not pruned) | (50% smearing, pruned) | (modified smearing with p = 0.5, not pruned) | (modified smearing with p = 0.5, pruned) | (modified smearing with p = 1, not pruned) | (modified smearing with p = 1, pruned) | (modified smearing with p = 2, not pruned) | (modified smearing with p = 2, pruned) |
| Using last week of Aug 2002 to predict 1st wk of Sep 2002 | Root Mean Square Error | 40.77929 | 40.94638 | 37.54829 | 37.85197 | 39.1553 | 39.2452 | 40.52794 | 40.58294 |
| | Mean Absolute Error | 28.53668 | 28.66426 | 26.72926 | 26.94496 | 27.82388 | 27.95006 | 28.46229 | 28.65281 |
| | Std. Deviation of Abs. Error | 29.14191 | 29.25088 | 26.38082 | 26.59467 | 27.55982 | 27.56003 | 28.86248 | 28.75096 |
| Using last week of Sep 2002 to predict 1st wk of Oct 2002 | Root Mean Square Error | 69.41525 | 69.53204 | 70.79193 | 71.07181 | 69.46349 | 69.79886 | 69.92595 | 70.16682 |
| | Mean Absolute Error | 43.89225 | 44.09121 | 44.81895 | 44.99941 | 43.73547 | 43.93878 | 45.11562 | 45.14456 |
| | Std. Deviation of Abs. Error | 53.78611 | 53.77424 | 54.80691 | 55.02092 | 53.97584 | 54.2428 | 53.43413 | 53.72475 |
| Using last week of Oct 2002 to predict 1st wk of Nov 2002 | Root Mean Square Error | 53.65864 | 53.75014 | 54.94735 | 55.02297 | 53.77958 | 53.95523 | 55.10207 | 55.47436 |
| | Mean Absolute Error | 39.68634 | 39.77412 | 40.15709 | 40.27475 | 39.17491 | 39.3886 | 40.30198 | 40.57747 |
| | Std. Deviation of Abs. Error | 36.12557 | 36.16506 | 37.5166 | 37.50129 | 36.85669 | 36.88565 | 37.58813 | 37.83867 |
| Using last week of Nov 2002 to predict 1st wk of Dec 2002 | Root Mean Square Error | 42.97271 | 43.07623 | 43.63672 | 43.63281 | 44.14822 | 44.26322 | 43.52 | 43.62062 |
| | Mean Absolute Error | 29.20839 | 29.33403 | 29.51306 | 29.43977 | 29.1324 | 29.20404 | 29.11515 | 29.32562 |
| | Std. Deviation of Abs. Error | 31.53281 | 31.55743 | 32.15539 | 32.21725 | 33.18507 | 33.27523 | 32.35946 | 32.3048 |

On the following pages we shall compare our hybrid approach with the earlier modelling methods. As the hybrid approach is a combination of neural networks (GRNN) and regression tree, its implementation also utilises the software packages used in our previous chapters, namely: "NeuralTools" and "rpart".

Specifically, we use NeuralTools to first construct a GRNN model from our data in an initial period. Then we write a simple Visual Basic script on the initial data to implement the "smearing" technique to generate a large set of artificial inputs (wool characteristics). This set of artificial inputs is then fed through the GRNN model to obtain a set of outputs (wool prices) predicted by the GRNN. The set of artificial inputs (wool characteristics), together with the outputs (wool prices) predicted by the GRNN, now form a complete set of artificial data (wool characteristics + prices). We can now use the rpart package to construct a single regression tree from the complete artificial set. And we shall use this regression tree to predict wool prices for the next period.

The tables on the following pages compare our hybrid approach with the earlier modelling methods. Our hybrid approach matches the accuracies of those of neural networks closely, and clearly a major leap from the more conventional single regression trees. This is a very attractive trade-off in getting a single tree representation considering a single tree is easy for causal observers to interpret and understand, as demonstrated in Chapter 4.

And of course the representer trees generated from our hybrid approach can also be expressed in the tabular form described in Section 4.4. After growing representer trees using our new hybrid approach, we can now come up with tabular representations of these trees using the method in Section 4.4, which streamline the process of assembling wool into bins and assist in delineating the specifications of individual bins.

Table 5.9: Fitting for Period A – Comparison of the Various Methods

| | | GRNN | Regression Tree (without pruning) | Bagging | Random Forrest | Hybrid Approach (modified smearing with p = 0.5, not pruned) |
|---|---|---|---|---|---|---|
| Fitting last week of Aug 2000 | Root Mean Square Error | 28.29091354 | 44.50163 | 50.39427 | 52.97568 | 28.85759 |
| | Mean Absolute Error | 15.26685036 | 23.3963 | 25.70137 | 28.17591 | 15.60698 |
| | Std. Deviation of Abs. Error | 23.82596829 | 37.86769 | 43.36212 | 44.87628 | 24.28116 |
| Fitting last week of Sep 2000 | Root Mean Square Error | 33.47677404 | 43.89331 | 50.58865 | 59.96171 | 33.95739 |
| | Mean Absolute Error | 16.33235456 | 23.13326 | 26.0236 | 31.0208 | 16.98274 |
| | Std. Deviation of Abs. Error | 29.23013271 | 37.31235 | 43.39332 | 51.32748 | 29.41342 |
| Fitting last week of Oct 2000 | Root Mean Square Error | 35.59787003 | 60.89278 | 64.65287 | 76.3059 | 36.31789 |
| | Mean Absolute Error | 20.77791974 | 29.00816 | 32.8266 | 37.83955 | 21.23019 |
| | Std. Deviation of Abs. Error | 28.9127874 | 53.55414 | 55.71469 | 66.28115 | 29.47456 |
| Fitting last week of Nov 2000 | Root Mean Square Error | 53.14404304 | 62.8968 | 82.35355 | 82.01819 | 53.43675 |
| | Mean Absolute Error | 17.53775253 | 30.5585 | 32.94694 | 40.16307 | 17.88617 |
| | Std. Deviation of Abs. Error | 50.1885246 | 54.99812 | 75.50842 | 71.54246 | 50.37617 |

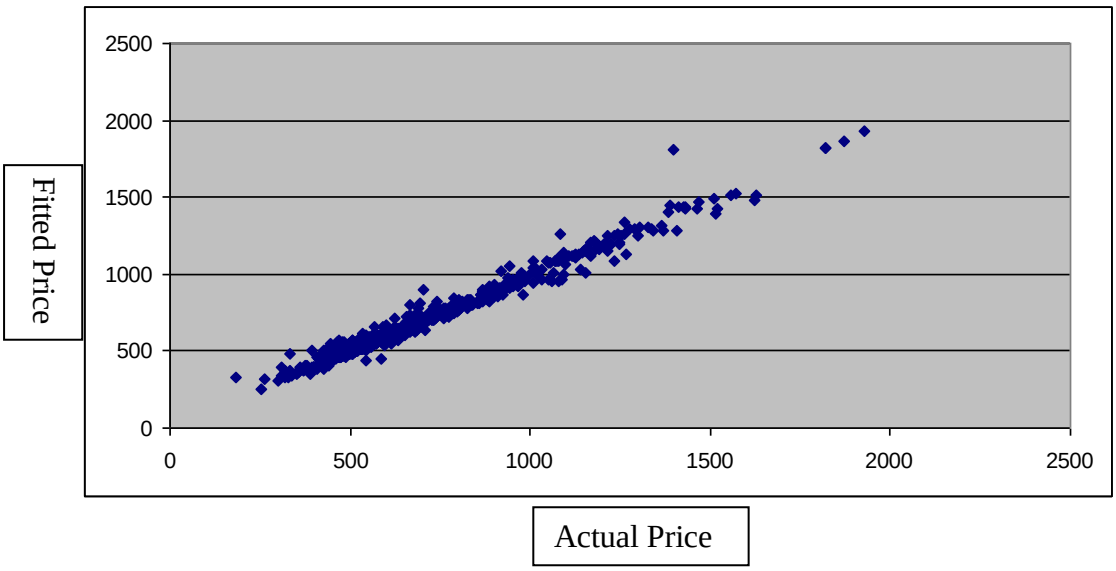


Figure 5.8: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of August 2000 with Hybrid Approach

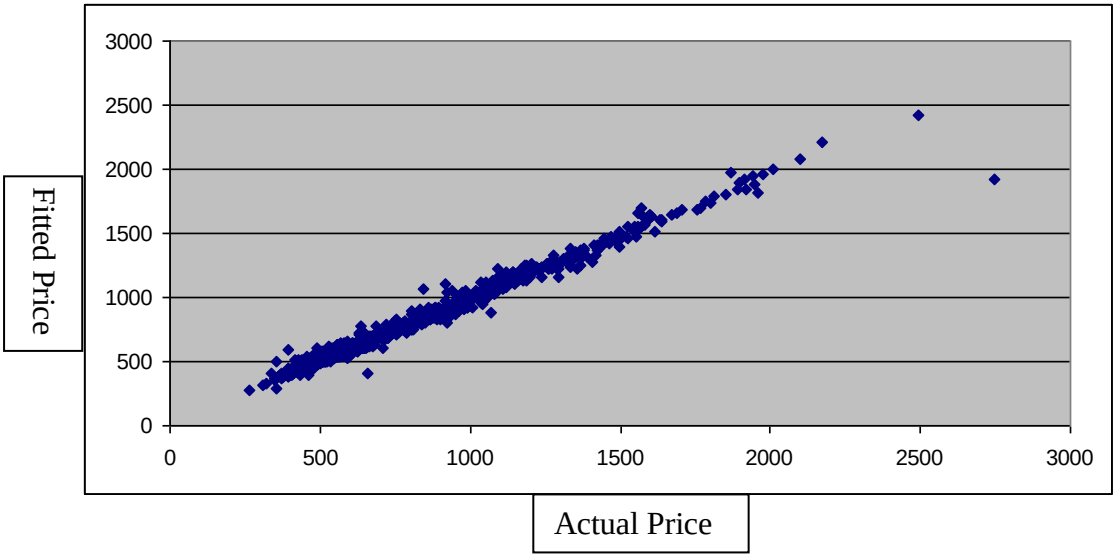


Figure 5.9: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of September 2000 with Hybrid Approach

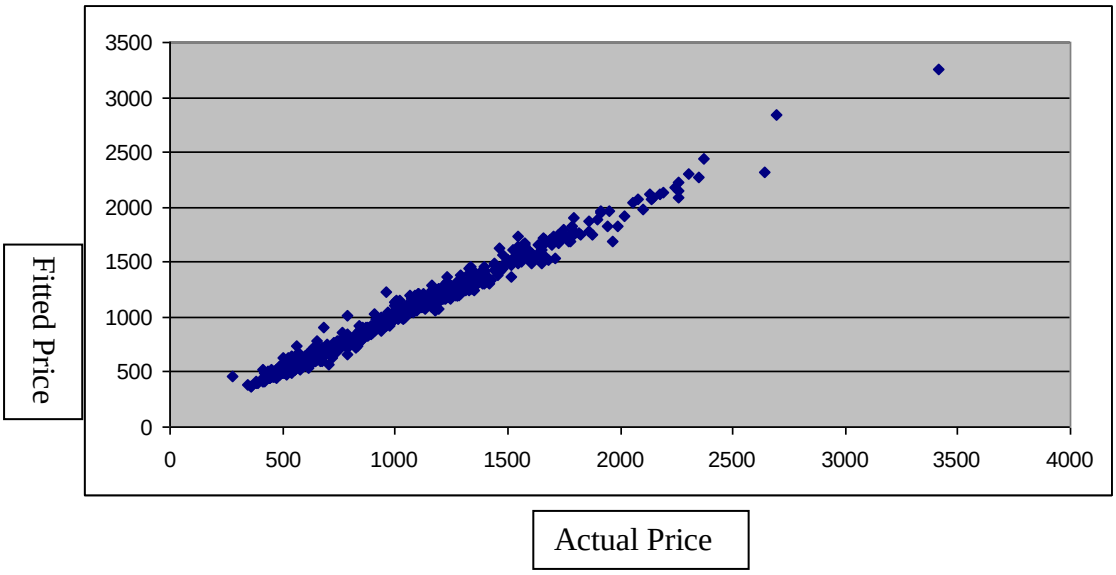


Figure 5.10: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of October 2000 with Hybrid Approach

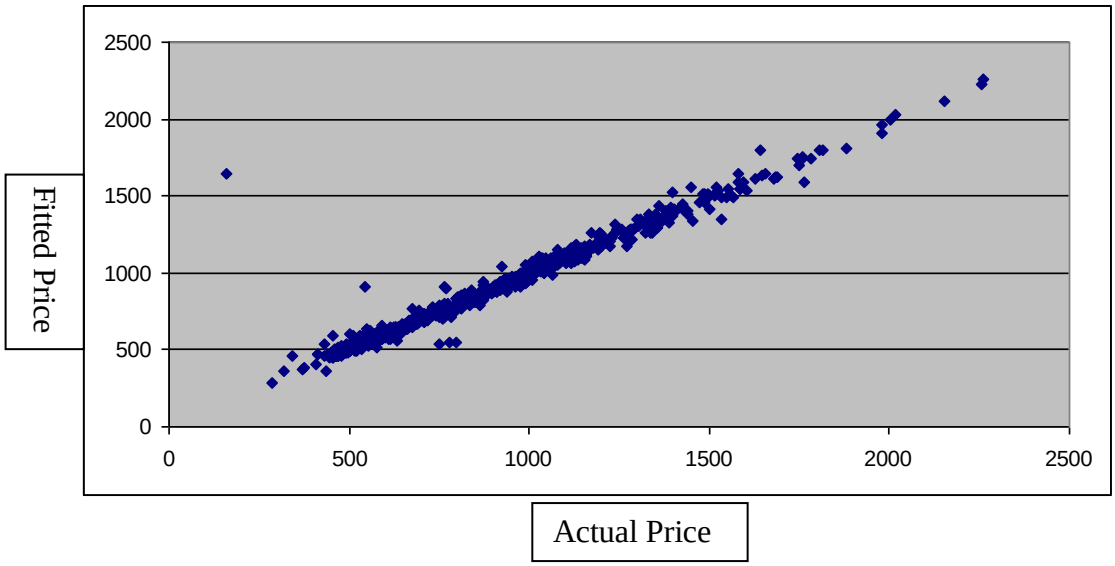


Figure 5.11: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of November 2000 with Hybrid Approach

Table 5.10: Fitting for Period B – Comparison of the Various Methods

| | | GRNN | Regression Tree (without pruning) | Bagging | Random Forrest | Hybrid Approach (modified smearing with p = 0.5, not pruned) |
|---|---|---|---|---|---|---|
| Fitting last week of Aug 2001 | Root Mean Square Error | 35.52785964 | 50.75957 | 58.01292 | 59.49253 | 36.11205 |
| | Mean Absolute Error | 19.05397627 | 27.20509 | 31.84648 | 33.15729 | 19.56654 |
| | Std. Deviation of Abs. Error | 29.99838122 | 42.87078 | 48.50984 | 49.4159 | 30.36407 |
| Fitting last week of Sep 2001 | Root Mean Square Error | 43.38224682 | 49.15919 | 58.10737 | 57.9892 | 44.28918 |
| | Mean Absolute Error | 18.91764899 | 21.53069 | 27.37857 | 27.32785 | 19.81226 |
| | Std. Deviation of Abs. Error | 39.04907165 | 44.20336 | 51.26467 | 51.15776 | 39.61962 |
| Fitting last week of Oct 2001 | Root Mean Square Error | 29.53830859 | 35.49992 | 44.13764 | 41.61895 | 29.78828 |
| | Mean Absolute Error | 16.2291678 | 19.29468 | 24.34452 | 23.20486 | 16.45713 |
| | Std. Deviation of Abs. Error | 24.6922348 | 29.81285 | 36.83433 | 34.56602 | 24.84133 |
| Fitting last week of Nov 2001 | Root Mean Square Error | 21.18334939 | 35.87093 | 42.29404 | 44.65671 | 21.57403 |
| | Mean Absolute Error | 11.16225224 | 19.22255 | 22.56911 | 22.9732 | 11.51276 |
| | Std. Deviation of Abs. Error | 18.01361591 | 30.30203 | 35.78841 | 38.31509 | 18.25531 |

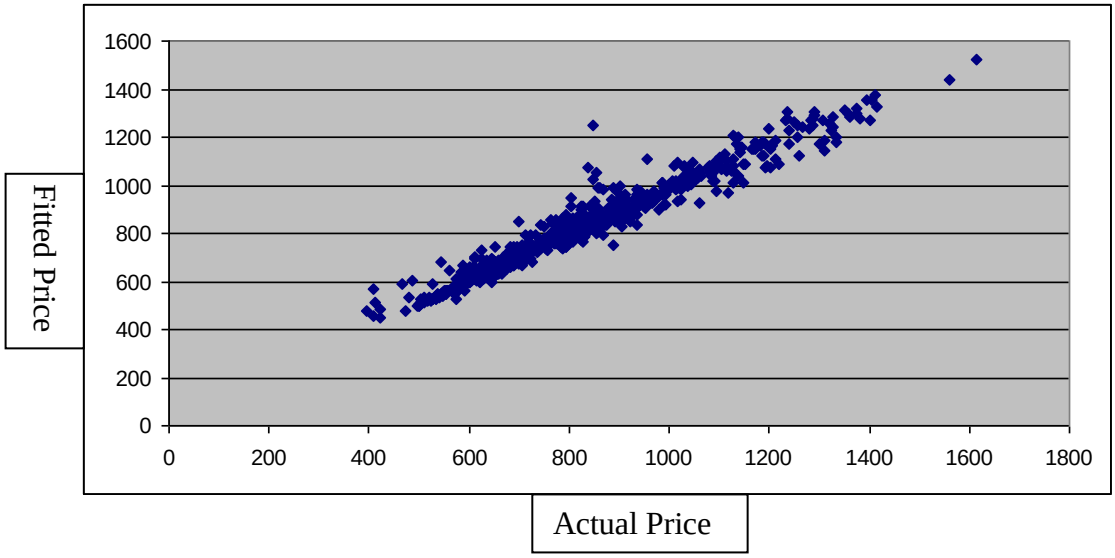


Figure 5.12: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of August 2001 with Hybrid Approach

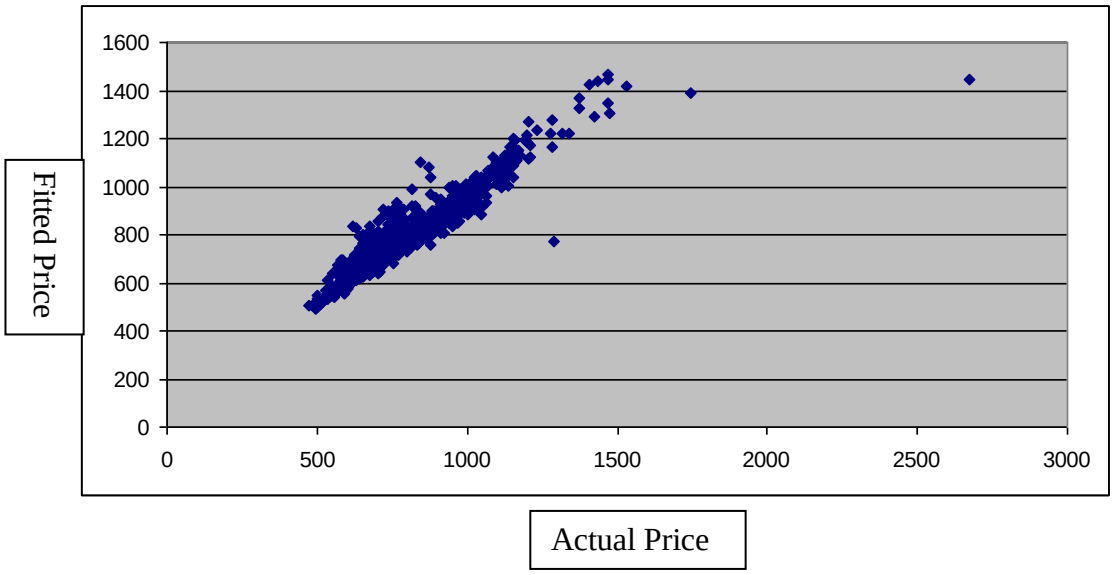


Figure 5.13: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of September 2001 with Hybrid Approach

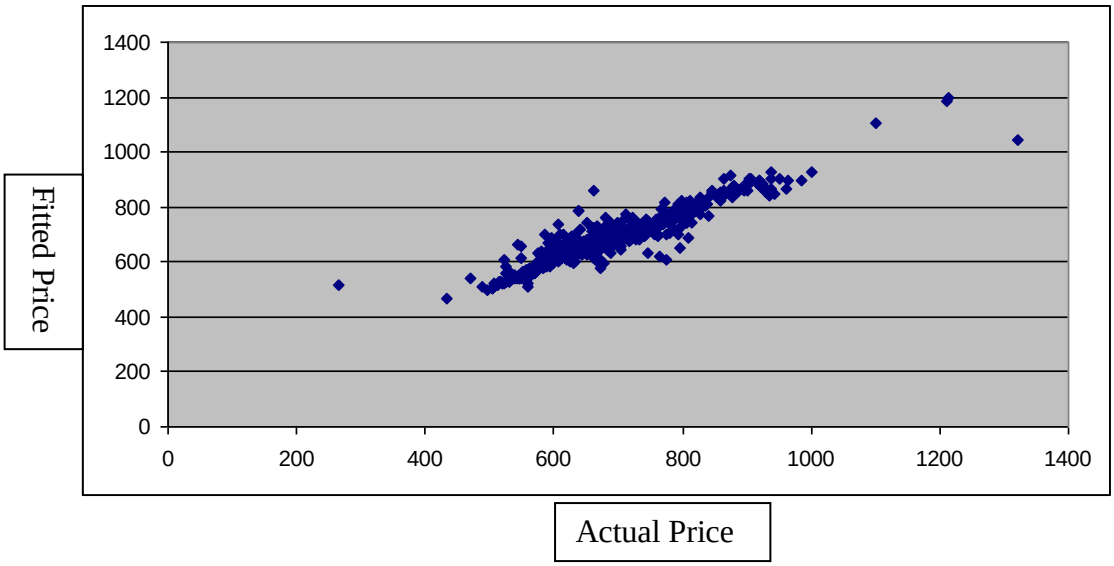


Figure 5.14: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of October 2001 with Hybrid Approach

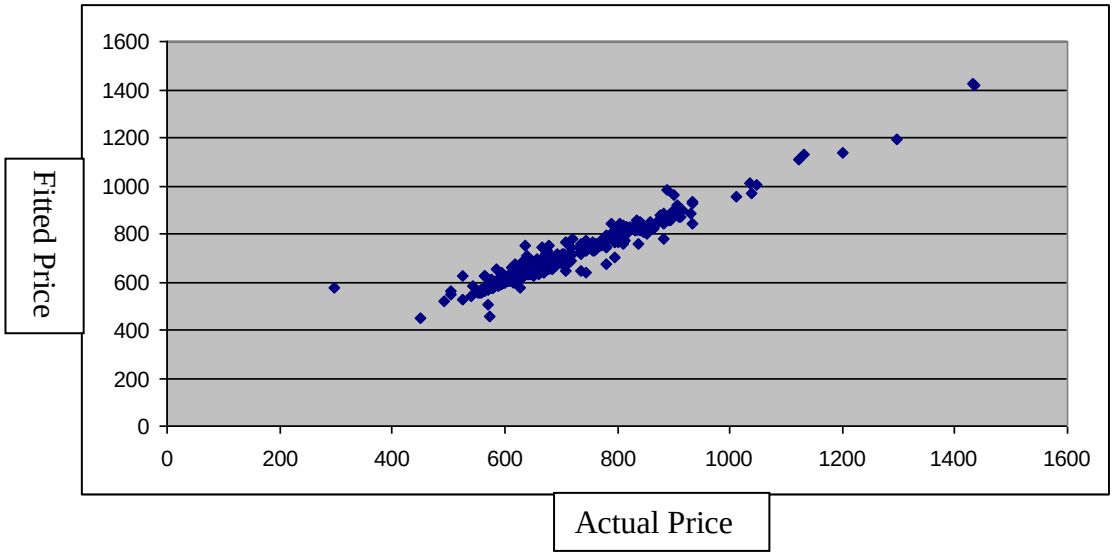


Figure 5.15: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of November 2001 with Hybrid Approach

Table 5.11: Fitting for Period C – Comparison of the Various Methods

| | | GRNN | Regression Tree (without pruning) | Bagging | Random Forrest | Hybrid Approach (modified smearing with p = 0.5, not pruned) |
|---|---|---|---|---|---|---|
| Fitting last week of Aug 2002 | Root Mean Square Error | 27.4876211 | 44.38823 | 54.82166 | 47.84726 | 28.32042 |
| | Mean Absolute Error | 15.47762742 | 23.93032 | 28.40097 | 26.17418 | 16.03741 |
| | Std. Deviation of Abs. Error | 22.72436108 | 37.39913 | 46.90882 | 40.06829 | 23.35065 |
| Fitting last week of Sep 2002 | Root Mean Square Error | 39.05447769 | 58.038 | 67.73973 | 65.07719 | 40.10445 |
| | Mean Absolute Error | 25.27669893 | 33.16222 | 40.58841 | 39.37019 | 26.16641 |
| | Std. Deviation of Abs. Error | 29.78166889 | 47.64693 | 54.25188 | 51.835 | 30.40262 |
| Fitting last week of Oct 2002 | Root Mean Square Error | 33.21652489 | 39.58191 | 48.19551 | 47.27497 | 33.79716 |
| | Mean Absolute Error | 19.23416188 | 24.98678 | 30.47481 | 29.50627 | 19.86391 |
| | Std. Deviation of Abs. Error | 27.09155177 | 30.71022 | 37.352 | 36.95077 | 27.35419 |
| Fitting last week of Nov 2002 | Root Mean Square Error | 33.74128347 | 38.34854 | 46.43959 | 44.55623 | 33.72553 |
| | Mean Absolute Error | 21.64586617 | 23.98833 | 28.34224 | 27.48117 | 21.79123 |
| | Std. Deviation of Abs. Error | 25.89736289 | 29.93598 | 36.80834 | 35.0914 | 25.75438 |

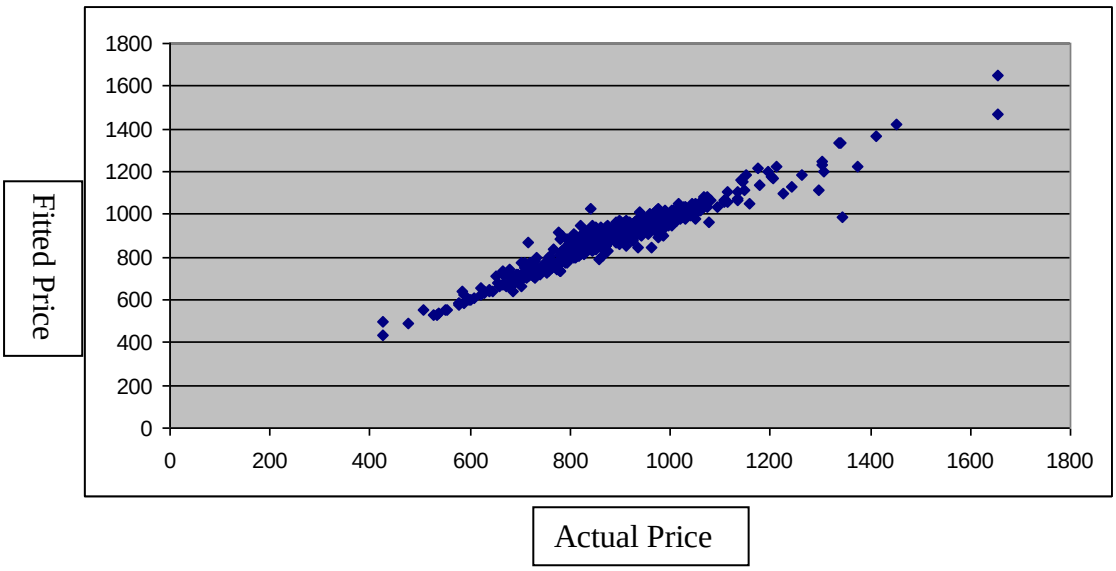


Figure 5.16: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of August 2002 with Hybrid Approach

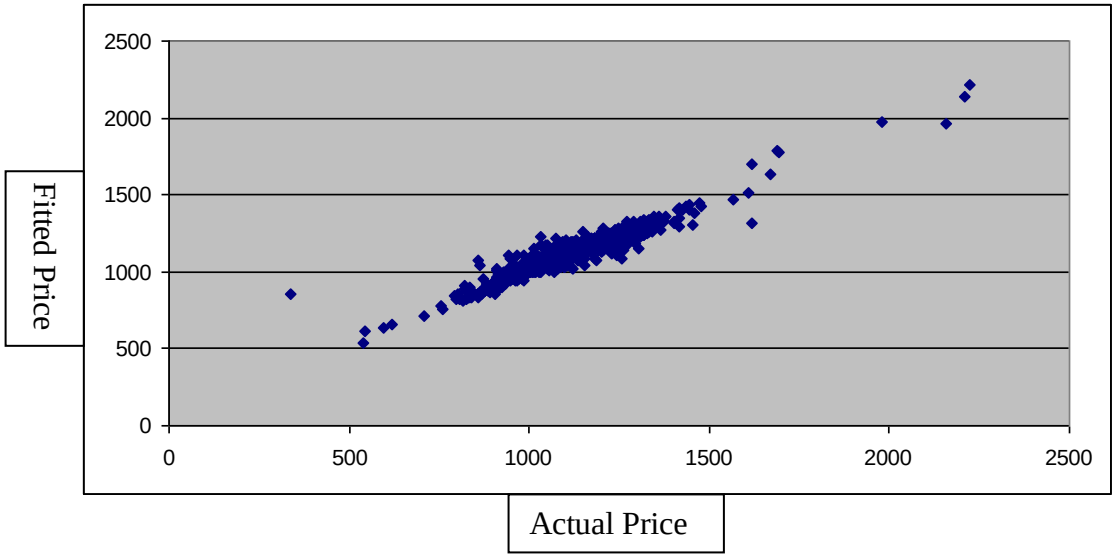


Figure 5.17: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of September 2002 with Hybrid Approach

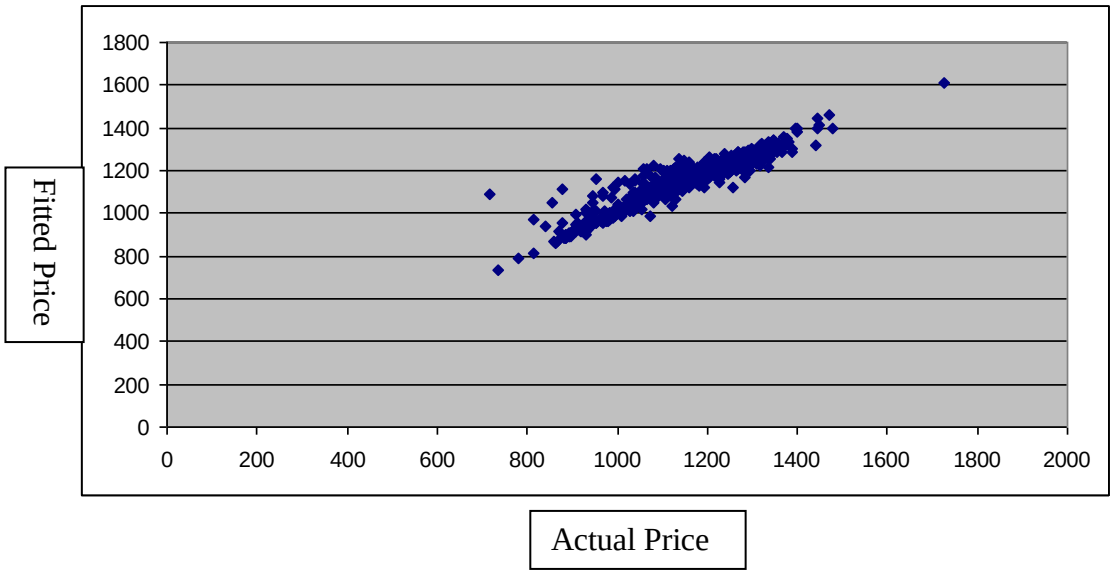


Figure 5.18: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of October 2002 with Hybrid Approach

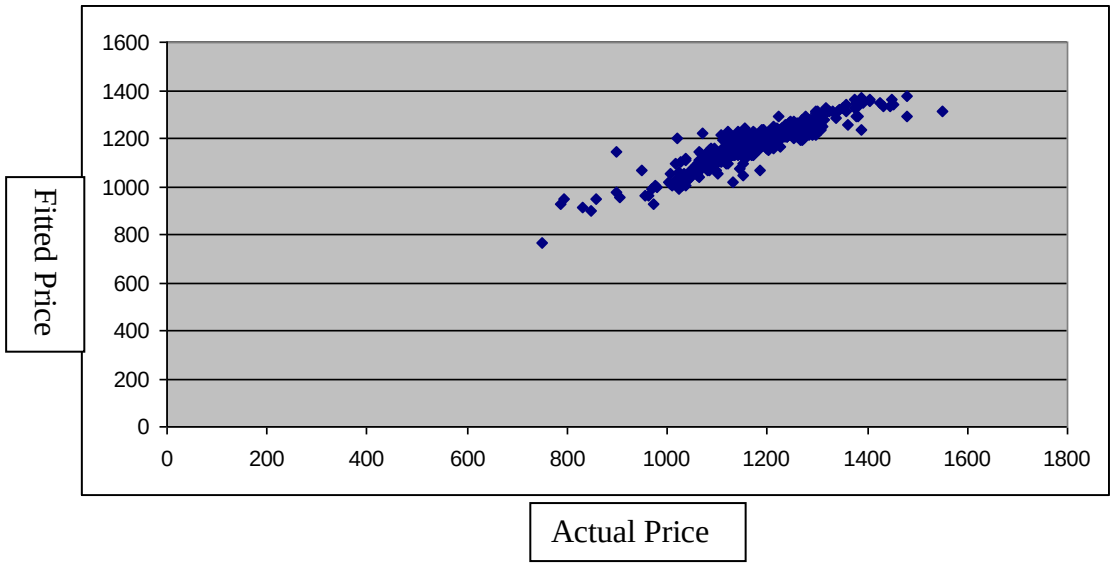


Figure 5.19: Fitted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of November 2002 with Hybrid Approach

Table 5.12: Predictions for Period A – Comparison of the Various Methods

| | | GRNN | Regression Tree (without pruning) | Bagging | Random Forrest | Hybrid Approach (modified smearing with p = 0.5, not pruned) |
|---|---|---|---|---|---|---|
| Using last week of Aug 2000 to predict 1st wk of Sep 2000 | Root Mean Square Error | 47.88411498 | 65.63687 | 50.9104 | 60.12618 | 51.4169 |
| | Mean Absolute Error | 27.49475198 | 34.04639 | 26.91575 | 30.97925 | 29.04302 |
| | Std. Deviation of Abs. Error | 39.21427868 | 56.13152 | 43.22526 | 51.54495 | 42.44027 |
| Using last week of Sep 2000 to predict 1st wk of Oct 2000 | Root Mean Square Error | 135.7182532 | 136.596 | 132.3349 | 148.8892 | 129.6546 |
| | Mean Absolute Error | 29.63332056 | 36.90987 | 29.8902 | 36.99417 | 29.77496 |
| | Std. Deviation of Abs. Error | 132.4945037 | 131.5653 | 128.9646 | 144.2755 | 126.2379 |
| Using last week of Oct 2000 to predict 1st wk of Nov 2000 | Root Mean Square Error | 69.12056444 | 96.05168 | 74.725 | 84.0915 | 74.0424 |
| | Mean Absolute Error | 35.6166384 | 48.66818 | 37.72461 | 45.26432 | 37.49623 |
| | Std. Deviation of Abs. Error | 59.2514565 | 82.82823 | 64.51828 | 70.88619 | 63.86079 |
| Using last week of Nov 2000 to predict 1st wk of Dec 2000 | Root Mean Square Error | 64.7849604 | 83.60047 | 69.56963 | 94.32711 | 61.89901 |
| | Mean Absolute Error | 34.43145383 | 45.3864 | 37.52524 | 44.5789 | 35.34225 |
| | Std. Deviation of Abs. Error | 54.90408942 | 70.24136 | 58.60961 | 83.16829 | 50.84185 |

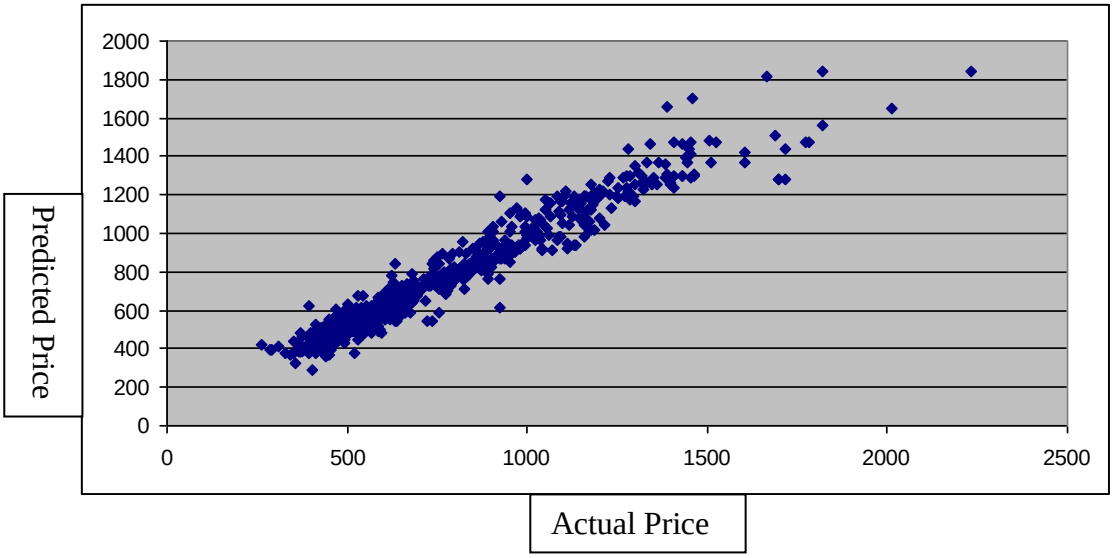


Figure 5.20: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of September 2000 with Hybrid Approach

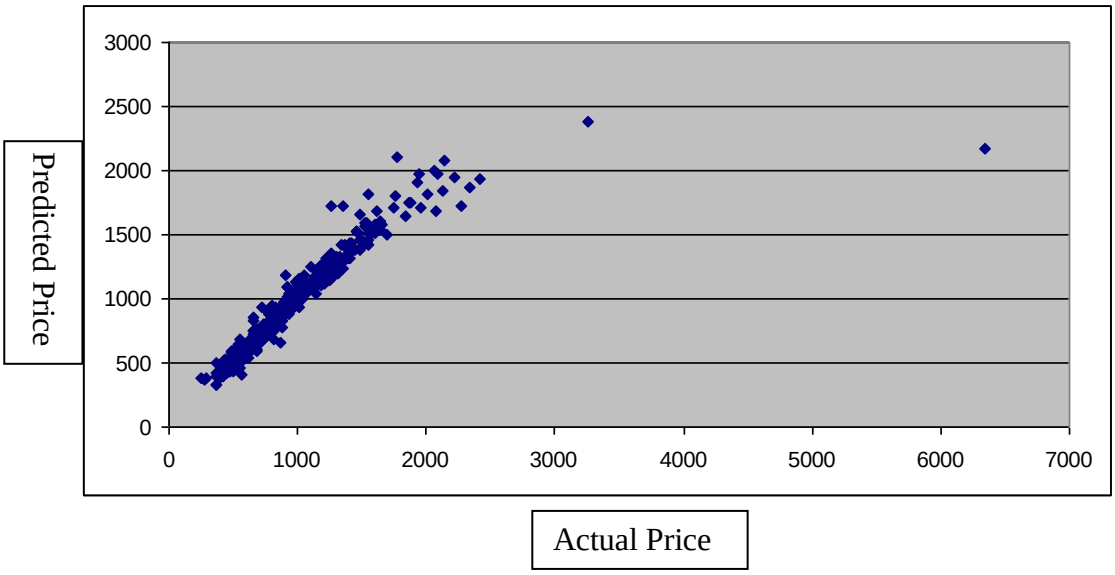


Figure 5.21: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of October 2000 with Hybrid Approach

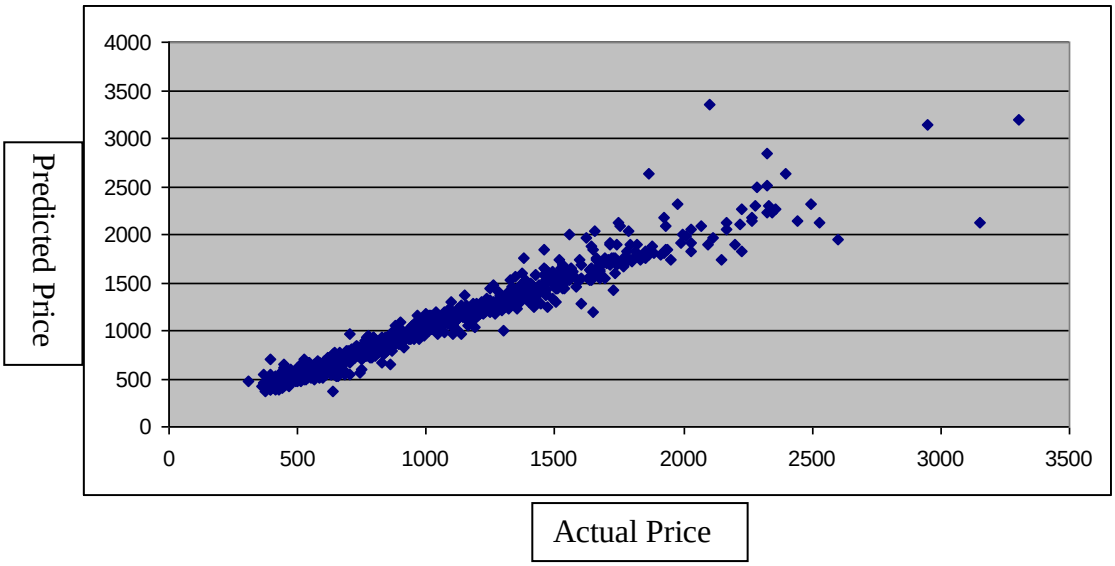


Figure 5.22: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of November 2000 with Hybrid Approach

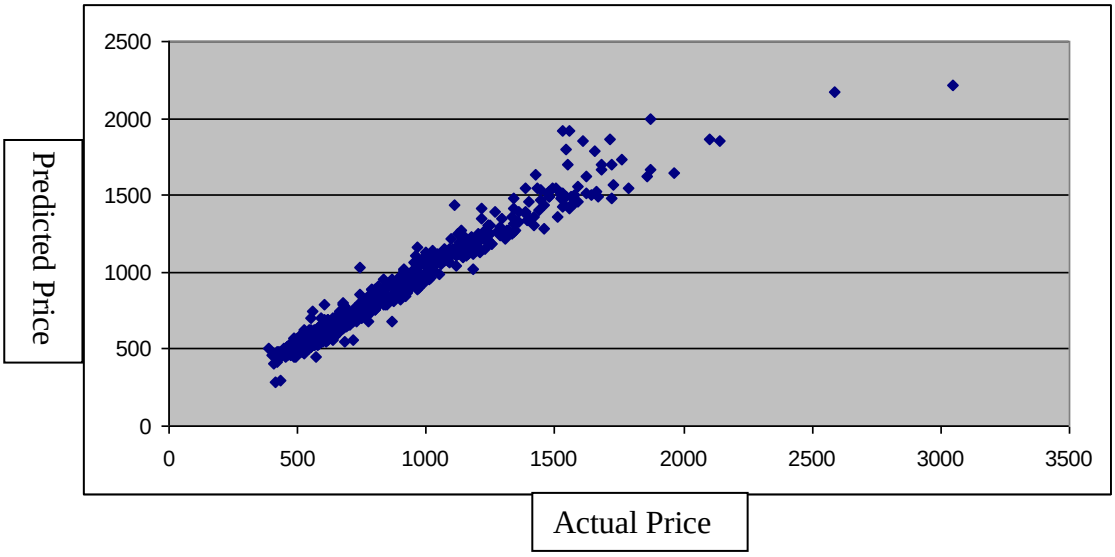


Figure 5.23: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of December 2000 with Hybrid Approach

Table 5.13: Predictions for Period B – Comparison of the Various Methods

| | | GRNN | Regression Tree (without pruning) | Bagging | Random Forrest | Hybrid Approach (modified smearing with p = 0.5, not pruned) |
|---|---|---|---|---|---|---|
| Using last week of Aug 2001 to predict 1st wk of Sep 2001 | Root Mean Square Error | 58.66584341 | 69.7189 | 58.85238 | 63.08733 | 60.49215 |
| | Mean Absolute Error | 32.43893849 | 36.92937 | 30.61869 | 32.26526 | 32.53699 |
| | Std. Deviation of Abs. Error | 48.8988455 | 59.15603 | 50.27818 | 54.23151 | 51.01466 |
| Using last week of Sep 2001 to predict 1st wk of Oct 2001 | Root Mean Square Error | 89.2013982 | 93.72725 | 88.41392 | 85.0301 | 88.2915 |
| | Mean Absolute Error | 74.55923056 | 73.82104 | 72.65839 | 72.55067 | 74.36393 |
| | Std. Deviation of Abs. Error | 48.97674904 | 57.76261 | 50.38595 | 44.35385 | 47.60521 |
| Using last week of Oct 2001 to predict 1st wk of Nov 2001 | Root Mean Square Error | 38.35475621 | 44.31331 | 39.72906 | 38.51702 | 38.57505 |
| | Mean Absolute Error | 27.41881763 | 30.44673 | 26.51407 | 26.55256 | 27.5051 |
| | Std. Deviation of Abs. Error | 26.82817574 | 32.20748 | 29.59656 | 27.91083 | 27.05489 |
| Using last week of Nov 2001 to predict 1st wk of Dec 2001 | Root Mean Square Error | 66.14964581 | 71.31696 | 66.97028 | 67.96612 | 67.10482 |
| | Mean Absolute Error | 54.72642547 | 57.22881 | 55.07718 | 52.90601 | 54.33174 |
| | Std. Deviation of Abs. Error | 37.18058425 | 42.5802 | 38.12095 | 42.69028 | 39.40709 |

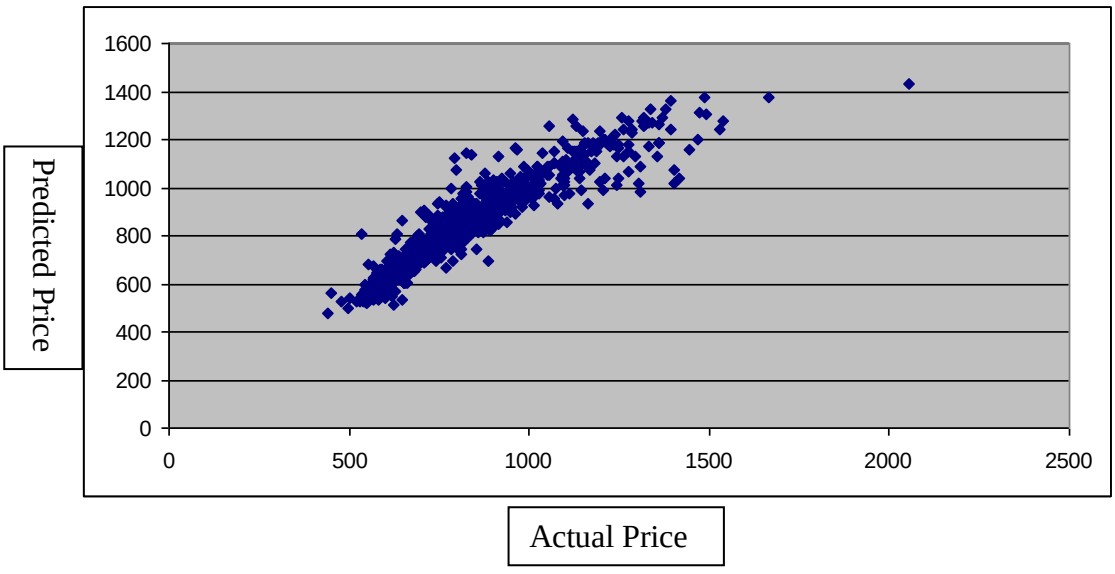


Figure 5.24: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of September 2001 with Hybrid Approach

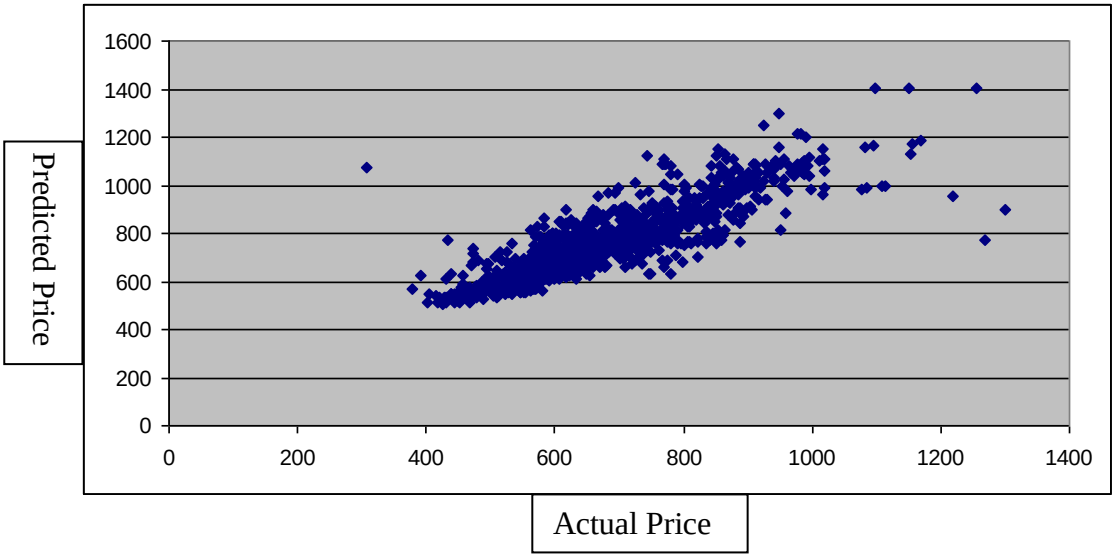


Figure 5.25: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of October 2001 with Hybrid Approach

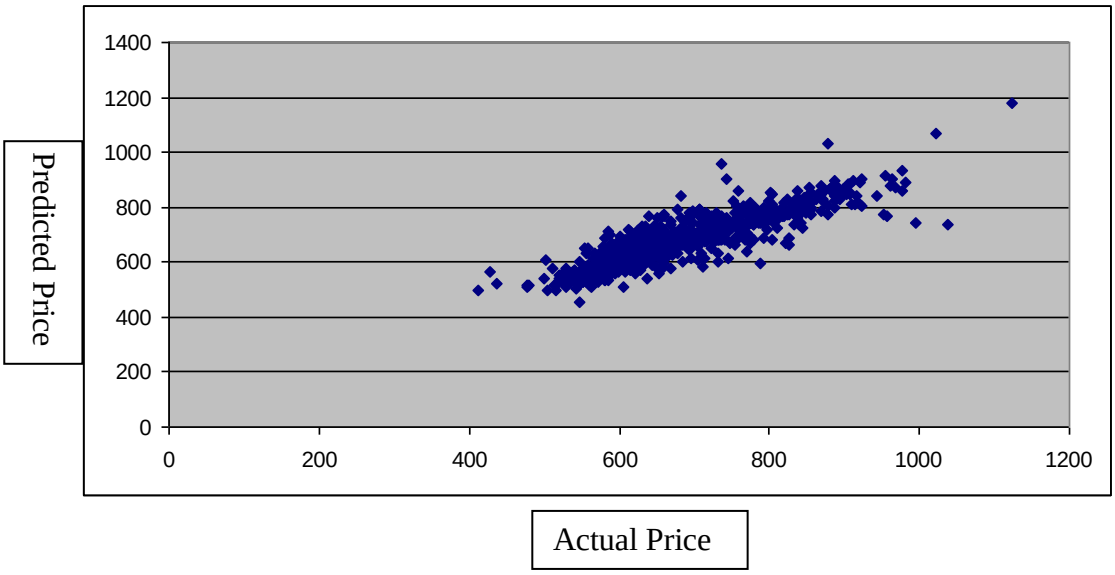


Figure 5.26: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of November 2001 with Hybrid Approach

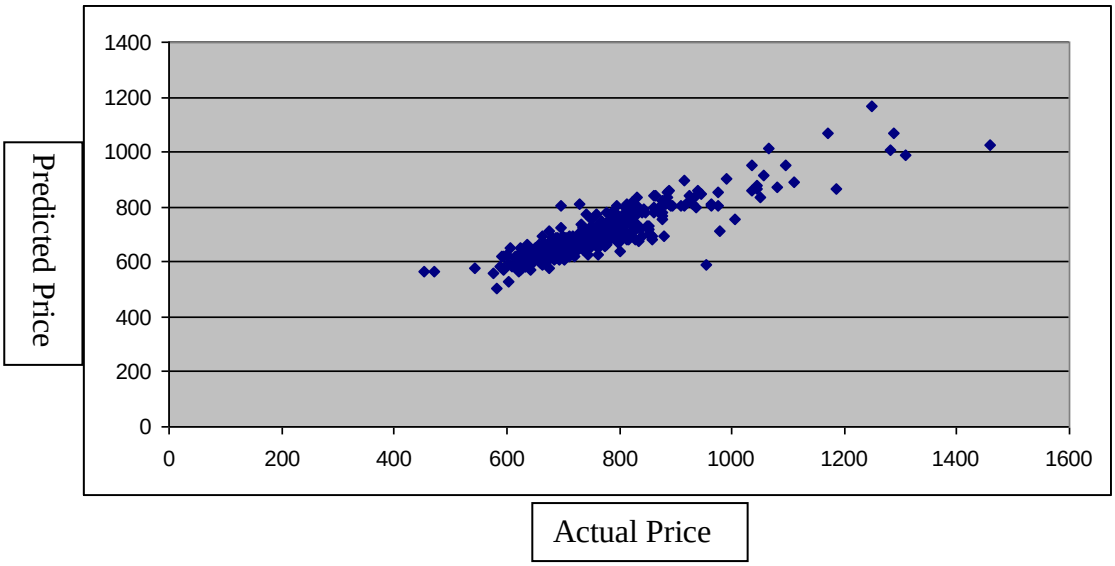


Figure 5.27: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of December 2001 with Hybrid Approach

Table 5.14: Predictions for Period C – Comparison of the Various Methods

| | | GRNN | Regression Tree (without pruning) | Bagging | Random Forrest | Hybrid Approach (modified smearing with p = 0.5, not pruned) |
|---|---|---|---|---|---|---|
| Using last week of Aug 2002 to predict 1st wk of Sep 2002 | Root Mean Square Error | 38.60712508 | 49.9932 | 38.81197 | 38.20277 | 37.54829 |
| | Mean Absolute Error | 27.09162225 | 32.91185 | 26.6451 | 26.55978 | 26.72926 |
| | Std. Deviation of Abs. Error | 27.51594552 | 37.64576 | 28.23138 | 27.47001 | 26.38082 |
| Using last week of Sep 2002 to predict 1st wk of Oct 2002 | Root Mean Square Error | 67.49464254 | 77.1083 | 66.68828 | 65.17472 | 70.79193 |
| | Mean Absolute Error | 42.78633291 | 49.20164 | 42.83662 | 42.84812 | 44.81895 |
| | Std. Deviation of Abs. Error | 52.20917584 | 59.38102 | 51.11999 | 49.11839 | 54.80691 |
| Using last week of Oct 2002 to predict 1st wk of Nov 2002 | Root Mean Square Error | 53.65067084 | 60.417 | 53.29528 | 50.8611 | 54.94735 |
| | Mean Absolute Error | 38.96359524 | 43.58813 | 39.29378 | 37.70949 | 40.15709 |
| | Std. Deviation of Abs. Error | 36.89281363 | 41.84948 | 36.01656 | 34.14046 | 37.5166 |
| Using last week of Nov 2002 to predict 1st wk of Dec 2002 | Root Mean Square Error | 41.68142591 | 46.99935 | 40.80626 | 41.99026 | 43.63672 |
| | Mean Absolute Error | 28.36965042 | 32.1573 | 28.09967 | 28.62148 | 29.51306 |
| | Std. Deviation of Abs. Error | 30.54913432 | 34.28974 | 29.60168 | 30.73675 | 32.15539 |

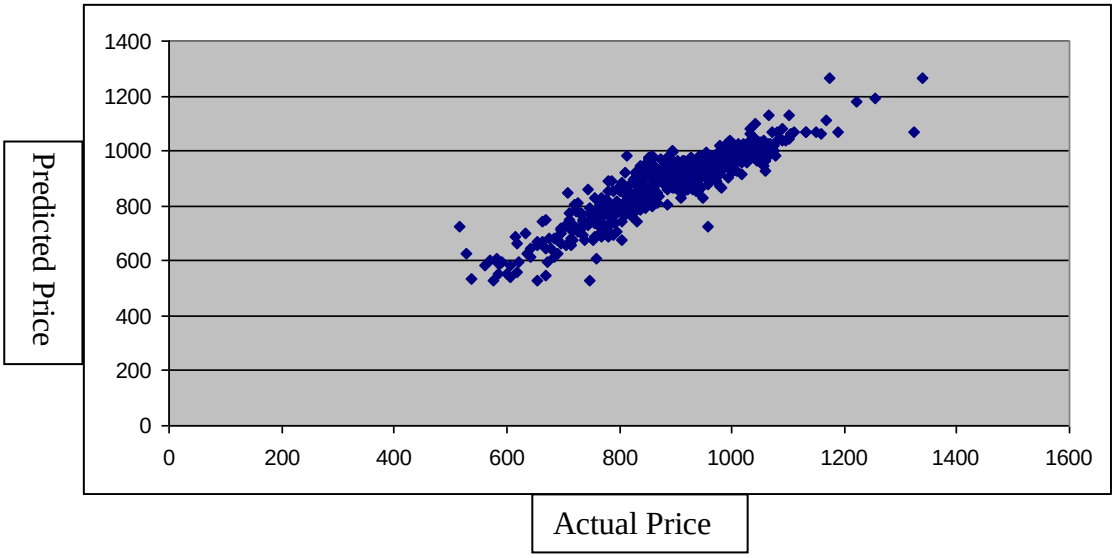


Figure 5.28: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of September 2002 with Hybrid Approach

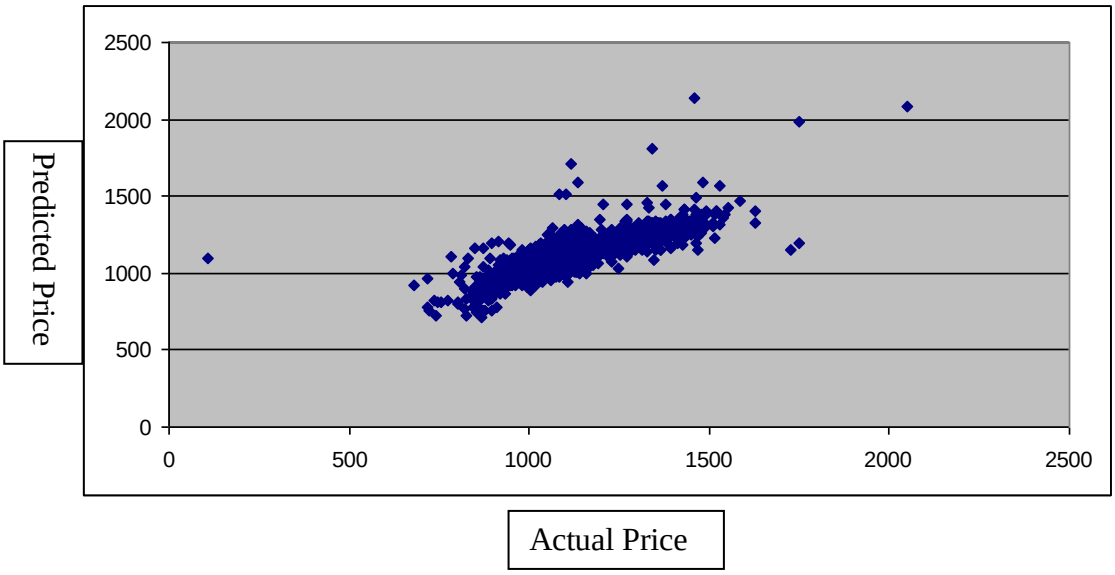


Figure 5.29: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of October 2002 with Hybrid Approach

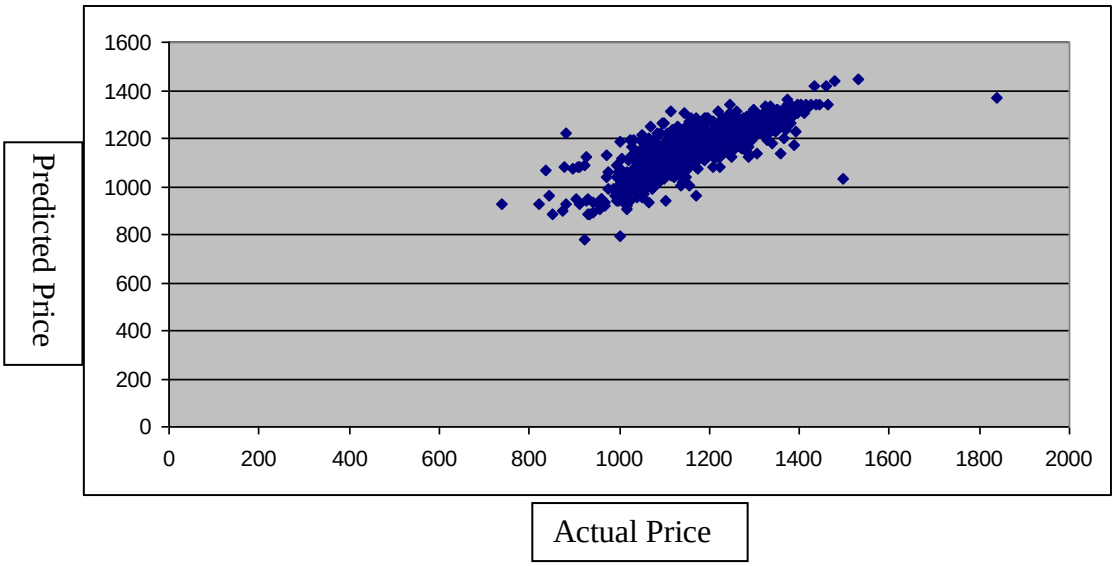


Figure 5.30: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of November 2002 with Hybrid Approach

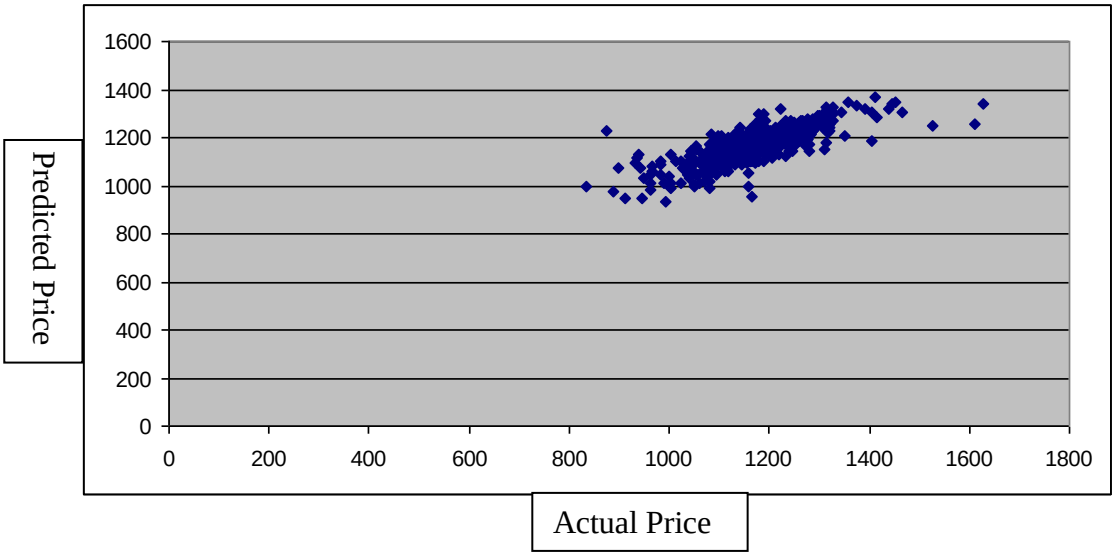


Figure 5.31: Predicted Price (vertical axis) vs. Actual Price (horizontal axis) in the last week of December 2002 with Hybrid Approach

## 5.4 Discussions on the Hybrid Approach

For fitting, tables 5.9 – 5.11 show that GRNN and the hybrid approach perform best for all periods, giving very similar results. In tables 5.12 – 5.14, the hybrid approach does not perform significantly better than other methods. This is not

unexpected, as the prediction component of our hybrid approach is built on the GRNN but with some errors/noise introduced when the "smearing" procedure is applied. It is thus expected that the hybrid approach would "mimic" the predictions and accuracy of GRNN, but never surpass it. The hybrid approach, however, still performs significantly better than an ordinary regression tree, given that both are the only methods that can generate a single tree required for interpretability. Given the handicaps, we consider that the hybrid approach still performs reasonably close to GRNN, with the slight decrease in accuracy a necessary trade-off.

It is debatable, if the phenomenon known as overfitting applies in our results. In our case, a data set to be predicted does not belong in the same week as the data set used for fitting. We only assume that the new week would share more or less very similar distributions and behaviours to the current week in our initial assumptions, but they should still be considered separate "populations".

In general, the hybrid approach detailed in this chapter has combined the best of both worlds: the accuracy of neural network methods and the superior interpretability of tree-based regression methods. The representer tree generated using this hybrid approach can also be expressed in tabular form in the same way describe in Section 4.4. This offers a very good solution to the wool specifications problem introduced in Section 1.3, and at the same time gives a much higher prediction accuracy than an ordinary regression tree.

The accuracy of our hybrid approach depends on three main factors:
(1) the data itself,
(2) the smearing method used, and
(3) the particular prediction model chosen as the function $\tilde{f}$.

There is not much we can do if the data given is rather limited. However, we can certainly make further improvements with factors (2) and (3). If a new predictive tool more accurate than GRNN comes along one day, then we can use this to replace GRNN in our procedure. Such is the advantage of our hybrid approach, which is highly modular and can be continually improved.

# Chapter 6

# Conclusion and Future Work

## 6.1 Summary of the Thesis

In this thesis, we developed and applied a new hybrid modular approach to the wool auction problem. Firstly, we presented a brief overview of the Australian wool auction market and discussed motivation and significance of our research. We defined the predictive aspect of the modelling problem and presented the data that was available. We also introduced the assumptions that must be made in order to model the auction data and predict the wool prices. We then examined neural networks and the family of tree-based regression methods and compared them as options for modelling and predicting the Australian wool auction prices.

We observed that neural network methods offered good prediction accuracy of price but gave minimal understanding of the price driving variables. On the other hand, tree-based regression methods offered good interpretability of the price driving characteristics but did not give good prediction accuracy of price. This motivated our hybrid approach that combined the best of the tree-based methods and neural networks, offering both prediction accuracy and interpretability.

Additionally, there also exists the wool specifications problem described in Section 1.3. Industrial sorting of wool during harvest, and at the start of processing, assembles wool in bins according to the required wool specifications. At present this assembly is done by constraining the range of all specifications in each bin, and having either a very large number of bins, or a large variance of characteristics within each bin. After growing representer trees using our hybrid approach developed in Chapter 5, we can now come up with tabular representations of these trees using the method in Section 4.4, which streamline

the process of assembling wool into bins and assist in delineating the specifications of individual bins.

Before this thesis, such methodology has not previously been used for wool auction data and the accompanying price prediction problem. Not only are the numeric predictions from our method comparable to other methods, our method also provides a clearer and better picture than ever before for a wool grower and other casual observers who may not have a higher level understanding of modelling and mathematics. This method is also highly modular and can be continually extended and improved.

## 6.2 Current Issues, Suggestions and Future Work

As mentioned at the end of Chapter 5, the accuracy of our hybrid approach depends on three main factors:
(1) the data itself,
(2) the smearing method used, and
(3) the particular prediction model chosen as the function $\tilde{f}$.

There is not much we can do if the data given is rather limited. However, we can certainly make further improvements with (2) and (3). If a new predictive tool more accurate than GRNN comes along one day, then we can use this to replace GRNN in our procedure. Such is the advantage of our hybrid approach, which is highly modular and can be continually improved.

Therefore, future work should concentrate on finding a modelling method/prediction tool that works well with our wool data and gives better predictions than GRNN (Goulermas et al. 2007), with the method's own interpretability no longer an important issue for consideration. This is because we are now able to integrate any such method into our hybrid modular approach and we can gain interpretability in the form of trees and tables from any method.

Besides regression tree, there also exist other more modern nonparametric regression techniques such as kernel regression and nonparametric multiplicative regression (NPMR) which also make very few model assumptions. Kernel regression estimates the continuous dependent variable from a limited set of data points by convolving the data points' locations with a kernel function, while NPMR is a smoothing technique based on multiplicative kernel estimation that can be cross-validated and applied in a predictive way. Both techniques are worth examining as future work on the wool auction problem.

We should also be able to improve the results in our hybrid approach further in another direction. The smearing procedure in our Section 5.3 was implemented using some coding in a copy of Excel 2003 that was available for us to use. The limitation of Excel 2003 is that each data file can only handle 65536 rows of data at a time, so we only generated up to 65535 rows of artificial data (sans heading) each time we performed our hybrid approach. This limitation has since been resolved with the release of Excel 2007 and Excel 2010. It is worth having a more detailed investigation into the effect of increasing the amount of artificial data use and at what level will the improvement to the results stop. Also, with more artificial data available, we will be able to have an even more in-depth look at the different variations of smearing and their differences in producing results at a much larger scale.

Another investigation worth looking into would be the possible integration of a time series component into our hybrid approach. Such a component was considered but not added as part of our project due to time and resource constraints. However, some preliminary ideas already exist and it would be interesting to investigate them.

And as briefly mentioned in Section 2.2, auction prices, like share and oil prices, depend not only on the product specifications and historical behaviour but also on intangible factors such as speculations, international market influences, and unexpected social and political events. To make our predictions as accurate as possible, ideally the intangible factors should be identified and captured in our

models, and their influences analysed. Considerations of the intangible factors will certainly strengthen our models and should be considered in future research.

# References

Allen, P.G. (1994), Economic forecasting in agriculture, *International Journal of Forecasting*, 10, 81–135.

Barutcuoglu Z. and E. Alpaydin (2003), A Comparison of Model Aggregation Methods for Regression, Proc. of 13 Int. Conf. on Artificial Neural Networks (ICANN),
http://www.cmpe.boun.edu.tr/~ethem/files/papers/27140076.pdf

Bessler, D.A. (1994), Economic forecasting in agriculture: Discussion, *International Journal of Forecasting*, 10, 137–138.

Breiman, L. (1996), Bagging Predictors, Machine Learning, 26(2), 123-140.

Breiman, L. (1996), Out-of-bag estimation,
ftp.stat.berkeley.edu/pub/users/breiman/OOBestimation.ps

Breiman, L. (1998), Arcing classifiers (discussion paper), *Annals of Statistics*, 26, 801-824.

Breiman, L. (1998), Randomizing outputs to increase prediction accuracy, Technical Report 518, May 1, 1998, Berkeley, CA: Statistics Department, University of California at Berkeley (in press, Machine Learning).

Breiman, L. (1999), Using adaptive bagging to debias regressions, Technical Report 547, Berkeley, CA: Statistics Department, University of California at Berkeley.

Breiman, L. (2000), Some infinity theory for predictor ensembles. Technical Report 579, Berkeley, CA: Statistics Department, University of California at Berkeley.

Breiman, L. (2001), Random forests, *Machine Learning*, 45(1), 5-32.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone (1984), Classification and Regression Trees, Wadworth and Brooks/Cole, Monterey.

Breiman, L., and N. Shang (1997), Born again trees, Technical Report, Berkeley, CA: Statistics Department, University of California at Berkeley.
ftp://ftp.stat.berkeley.edu/pub/users/breiman/BAtrees.ps

Caccetta L., C. Chow, T. Dixon, and J. Stanton (2005), Modelling the Structure of Australian Wool Auction Prices, Conference Proceedings, International Congress on Modelling and Simulation (MODSIM05), (Editors: A. Zerger and R.M. Argent), Melbourne, Australia, December, 2005.
http://www.mssanz.org.au/modsim05/papers/caccetta_1.pdf

Caccetta L., C.N. Chow, K. Curtis, and J. Stanton (2007), Modelling the Structure of Australian Wool Auction Prices: A Hybrid Approach Combining Regression Tree and Neural Networks, Proceedings, The 7th International Conference on Optimization: Techniques and Applications (ICOTA7), (Editors: Masao Fukushima et al.), Kobe, Japan, December, 2007. ISBN 978-4-946443-15-2.

Caccetta L., C.N. Chow, and J. Stanton (2009), Modelling the Structure of Australian Wool Auction Prices: A Hybrid Approach Combining Regression Tree and Neural Networks, Conference Proceedings, The 20th National Conference of Australian Society for Operations Research (ASOR Conference 2009), Gold Coast, Australia, September, 2009.

Cheng B. and D.M. Titterington (1994), Neural Networks: A Review from a Statistical Perspective, Statistical Science, Vol. 9, No. 1. (Feb., 1994), pp. 2-30.
http://links.jstor.org/sici?sici=0883-4237%28199402%299%3A1%3C2%3ANNARFA%3E2.0.CO%3B2-A

Cheng, Y.W., J. Stanton, and L. Caccetta, (2004), Predicting the Australian wool auction price by tree-based regression, in Proceeding of Industrial Optimisation Symposium, Curtin University of Technology, Western Australia.

Craven, M., and W. Shavlik (1996), Extracting tree-structured representations of trained networks, Advances in Neural Information Processing Systems, 8, 24-30.
Freebairn, J. (1994), The agricultural commodity market forecasting game, *International Journal of Forecasting*, 10, 139–142.

Frank E. and B. Pfahringer (2006), Improving on Bagging with Input Smearing, *Advances in Knowledge Discovery and Data Mining*, *Lecture Notes in Computer Science*, 2006, Volume 3918/2006, 97-106, DOI: 10.1007/11731139_14, http://www.cs.waikato.ac.nz/~eibe/pubs/FrankAndPfahringer.pdf

Freebairn, J. (1994), The agricultural commodity market forecasting game, International Journal of Forecasting, 10, 139–142.

Freund Y. and R.E. Schapire (1995), A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, http://dnkweb.denken.or.jp/boosting/papers/FreSch97.ps.gz

Friedman, J.H. (2001), Greedy Function Approximation: A Gradient Boosting Machine, *The Annals of Statistics*, Vol. 29, No. 5, pp. 1189-1232. http://links.jstor.org/sici?sici=0090-5364%28200110%2929%3A5%3C1189%3AGFAAGB%3E2.0.CO%B2-N

Friedman J., T. Hastie and R. Tibshirani (2000), Additive Logistic Regression: a Statistical View of Boosting, *The Annals of Statistics*, Vol. 28, No. 2, pp. 337–407.

Frost, F. and V. Karri (1999), Performance Comparison of BP and GRNN Models of the Neural Network Paradigm Using a Practical Industrial Application, Proc. 6th International Conference on Neural Information Processing (ICONIP), Nov. 1999, Perth., pp. 1069-1075.

Goulermas J.Y., P. Liatsis, X.J. Zeng and P. Cook (2007), Density-Driven Generalized Regression Neural Networks (DD-GRNN) for Function Approximation, *IEEE Transactions on Neural Networks*, Vol. 18, No. 6, November 2007.

Goulermas J.Y., X.J. Zeng, P. Liatsis and J.F. Ralph (2007), Generalized Regression Neural Networks With Multiple-Bandwidth Sharing and Hybrid Optimization, *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, Vol. 37, No. 6, December 2007.

Graham-Higgs, J., A. Rambaldi, and B. Davidson (1999), Is the Australian wool futures market efficient as a predictor of spot prices?, *Journal of Futures Markets*, 19(5), 565–582.

Ho, T.K. (1995). Random Decision Forests. 3rd Int'l Conf. on Document Analysis and Recognition. pp. 278–282.
http://cm.bell-labs.com/cm/cs/who/tkh/papers/odt.pdf

Jones, C., F. Menezes, and F. Vella (2004), Auction price anomalies: evidence from wool auctions in Australia, *The Economic Record*, 80(250), 271–288.

Kemp, S., and K. Willetts (1996), Remembering the price of wool, *Journal of Economic Psychology*, 17, 115–125.

Morgan, J.N., and R.C. Messenger (1973), *THAID: A sequential search program for the analysis of nominal scale dependent variables*, Technical report, Survey Research Center, Institute for Social Research, University of Michigan, Michigan.

Morgan, J.N., and J.A. Sonquist (1963), Problems in the analysis of survey data, and a proposal, *Journal of the American Statistical Association*, 58, 415–434.

Quinlan, J.R. (1979), Discovering rules by induction from collections of examples, in *Expert Systems in the Microelectronic Age*, ed. D. Michie, Edinburgh University Press, Edinburgh.

Quinlan, J.R. (1983), Learning efficient classification procedures and their application to chess end-games, in *Machine Learning*, eds R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, Tioga, 463–482, Palo Alto.

Quinlan, J.R. (1986), Induction of decision trees, *Machine Learning*, 1, 81–106.
Quinlan, J.R. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.

Scott, C.D., R.M. Willett, and R.D. Nowak (2003), CORT: Classification or regression trees, in *Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Simmons, P., and P. Hansen (1997), The effect of buyer concentration on prices in the Australian wool market, *Agribusiness*, 13(4), 423–430.

Specht, D.F. (1991), A general regression neural network, IEEE Transactions on Neural Networks, Vol. 2, No. 6, November 1991.

Stanton, J.H. (1993), Analyses of auction prices from Fremantle and their comparison with Eastern State prices, Wool Processing Research Opportunities, Department of Agriculture WA.

Stanton, J.H. (1994), Western Australian wool production. Part 1: Analysis by weight and characteristics, Department of Agriculture WA.

Stanton, J.H., and L.R. Coss (1995), Characteristics of wool from shires in the Northern Region, Rural Research for Farm Profit, Department of Agriculture WA, 157–158.

Stanton, J.H., K. Curtis, and L.R. Coss (1997), Application of auction information to wool processing, *IWTO Conference*, Boston.

Tomek, W.G. (1994), Economic forecasting in agriculture: Comment, *International Journal of Forecasting*, 10, 143–145.

Watters, G., and R. Deriso (2000), Catches per unit of effort of bigeye tuna: a new analysis with regression trees and simulated annealing, *Inter-American Tropical Tuna Commission*, 21(8), 531–571.